der Bundeswehr Universität München
der Bundeswehr Universität München
Research Center Cyber Defence Universität der Bundeswehr München
Research Center Cyber Defence Universität der Bundeswehr München

Proceedings of the
15th European Conference on Cyber Warfare and Security

Bundeswehr University
Munich, Germany
7-8 July 2016

Edited by

Gabi Rodosek and Robert Koch

A conference managed by ACPI, UK

acpi

Proceedings of The 15th European Conference on Cyber Warfare and Security

ECCWS 2016

Hosted by
Universität der Bundeswehr
Munich
Germany

7-8 July 2016

Edited by
Robert Koch and Gabi Rodosek
Bundeswehr University, Munich, Germany

Review Process
Papers submitted to this conference have been double-blind peer reviewed before final acceptance to the conference. Initially, abstracts were reviewed for relevance and accessibility and successful authors were invited to submit full papers. Many thanks to the reviewers who helped ensure the quality of all the submissions.

Ethics and Publication Malpractice Policy
ACPIL adheres to a strict ethics and publication malpractice policy for all publications – details of which can be found here:
http://www.academic-conferences.org/policies/ethics-policy-for-publishing-in-the-conference-proceedings-of-academic-conferences-and-publishing-international-limited/

Conference Proceedings
The Conference Proceedings is a book published with an ISBN and ISSN. The proceedings have been submitted to a number of accreditation, citation and indexing bodies including Thomson ISI Web of Science and Elsevier Scopus.

Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

The Electronic version of the Conference Proceedings is available to download from DROPBOX. (http://tinyurl.com/ECCWS2016) Select Download and then Direct Download to access the Pdf file. Free download is available for conference participants for a period of 2 weeks after the conference.

The Conference Proceedings for this year and previous years can be purchased from http://academic-bookshop.com

Print version ISSN: 2048-8610
Print version ISBN: 978-1-910810-93-4
E-Book ISSN: 2048-8602
E-Book ISBN: 978-1-910810-96-5

Published by Academic Conferences and Publishing International Limited
Reading, UK. 44-118-972-4148. www.academic-publishing.org

Contents

Preface

These proceedings represent the work of researchers participating in the 15th European Conference on Cyber Warfare and Security (ECCWS 2016) which is being hosted this year by the Universität der Bundeswehr, Munich, Germany on the 7-8 July 2016.

ECCWS is a recognised event on the International research conferences calendar and provides a valuable platform for individuals to present their research findings, display their work in progress and discuss conceptual and empirical advances in the area of Cyberwar and Cyber Security. It provides an important opportunity for researchers and managers to come together with peers to share their experiences of using the varied and expanding range of Cyberwar and Cyber Security research available to them.

The conference this year will be opened with a keynote presentation by Kai Horten, ESG Elektroniksystem und Logistik-GmbH, Germany. The second day of the conference will be opened by Dr Detlef Houdeau, Infineon Technologies, Germany. By alluring these speakers, the conference will bridge academic research and real-world cyber security requirements, pointing out the challenges of exploring adaptable and applicable systems and technologies.

With an initial submission of 110 abstracts, after the double blind, peer review process there are 37 Academic research papers and 11 PhD research papers, 1 Master's research paper, 2 Work In Progress papers and 2 non-academic papers published in these Conference Proceedings. These papers come from many different countries including Austria, Belgium, Canada, Czech Republic, Finland, France, Germany, Greece, Hungary, Ireland, Kenya, Luxembourg, Netherlands, Norway, Portugal, Romania, Russia, Slovenia, South Africa, Sweden, Turkey, UK and USA. This is not only highlighting the international character of the conference, but is also promising very interesting discussions based on the broad treasure trove of experience of our community and participants.

We wish you a most interesting conference, a lot of opportunities for networking and a boost for your research!

Robert Koch, Programme Chair
And
Gabi Rodosek, Conference Chair

July 2016

Conference Committee

Dr. Mohd Faizal Abdollah (University Technical Malaysia Melaka, Melaka); Dr. Nasser Abouzakhar (University of Hertfordshire, UK); Dr. Kari Alenius (University of Oulu, Finland); Chaminda Alocious (University of Hertfordshire, UK); Prof. Antonios Andreatos (Hellenic Air Force Academy, Greece); Dr. Olga Angelopoulou (University of Derby, UK); Dr. Leigh Armistead (Edith Cowan University, Australia); Colin Armstrong (Curtin University, Australia, Australia); Johnnes Arreymbi (University of East London, UK); Debi Ashenden (Cranfield University, Shrivenham, UK); Dr. Darya Bazarkina (Sholokhov Moscow State Humanitarian University, Russian Federation); Laurent Beaudoin (ESIEA, Laval, France); Ass Prof. Maumita Bhattacharya (Charles Sturt University, Australia); Prof. Matt Bishop (University of California at Davis, USA); Andrew Blyth (University of Glamorgan, UK); Colonel (ret) Colin Brand (Graduate School of Business Leadership, South Africa); Dr. Svet Braynov (University of Illinois at Springfield, USA); Prof. Larisa Breton (University of the District of Columbia, USA); Bill Buchanen (Napier University, UK); Dr. Joobin Choobineh (Texas A&M University, USA); Bruce Christianson (University of Hertfordshire, UK); Dr. Maura Conway (Dublin City University, Ireland); Dr. Paul Crocker (Universidade de Beira Interior, Portugal); Prof. Tiago Cruz (University of Coimbra, Portugal); Dr. Christian Czosseck (CERT Bundeswehr (German Armed Forces CERT), Germany); Geoffrey Darnton (Bournemouth University, UK); Josef Demergis (University of Macedonia, Greece); Dr. Martina Doolan (University of Hertfordshire, UK); Paul Dowland (University of Plymouth, UK); Marios Efthymiopoulos (Political Science Department University of Cyprus, Cyprus); Dr. Colin Egan (University of Hertfordshire, Hatfield, UK); Dr. Ramzi El-Haddadeh (Brunel University, UK); Daniel Eng (C-PISA/HTCIA, China); Prof. Dr. Alptekin Erkollar (ETCOP, Austria); Prof. Robert Erra (ESIEA PARIS, France); John Fawcett (University of Cambridge, UK); Prof. Eric Filiol (Ecole Supérieure en Informatique, Electronique et Automatique, France); Dr. Chris Flaherty (University of New South Wales, Australia); Prof. Steve Furnell (University of Plymouth, UK); Assoc. Professor Javier Garci'a Villalba (Universidad Complutense de Madrid, Spain); Kevin Gleason (KMG Consulting, MA, USA); Dr. Michael Grimaila (Air Force Institute of Technology, USA); Prof. Stefanos Gritzalis (University of the Aegean, Greece); Dr. Marja Harmanmaa (University of Helsinki, Finland); Ulrike Hugl (University of Innsbruck, Austria); Aki Huhtinen (National Defence College, Finland); Bill Hutchinson (Edith Cowan University, Australia); Dr. Berg Hyacinthe (State University of Haiti, Haiti); Dr. Abhaya Induruwa (Canterbury Christ Church University, UK); Hamid Jahankhani (University of East London, UK); Dr. Helge Janicke (De Montfort University, UK); Joey Jansen van Vuuren (CSIR, South Africa); Saara Jantunen (University of Helsinki, Finland); Andy Jones (BT, UK); Dr. Audun Josang (University of Oslo, Norway); James Joshi (University of Pittsburgh, USA); Nor Badrul Anuar Jumaat (University of Malaya, Malaysia); Maria Karyda (University of the Aegean, Greece); Ass Prof. Vasilis Katos (Democritus University of Thrace, Greece); Dr. Anthony Keane (Institute of Technology Blanchardstown, Dublin, Ireland); Jyri Kivimaa (Cooperative Cyber Defence and Centre of Excellence, Tallinn, Estonia); Dr. Spyros Kokolakis (University of the Aegean, Greece); Prof. Ahmet Koltuksuz (Yasar University, Dept. of Comp. Eng., Turkey); Theodoros Kostis (Hellenic Army Academy, Greece); Prashant Krishnamurthy (University of Pittsburgh, USA); Dan Kuehl (National Defense University, Washington DC, USA); Peter Kunz (Diamler, Germany); Pertti Kuokkanen (Finnish Defence Forces, Finland); Dr. Erikk Kurkinen (University of Jyväskylä, Finland); Takakazu Kurokawa (National Defence Acadamy, Japan); Rauno Kuusisto (Finnish Defence Force, Finland); Tuija Kuusisto (National Defence University, Finland); Dr. Laouamer Lamri (Al Qassim University and European University of Brittany, Saudi Arabia); Michael Lavine (John Hopkins University's Information Security Institute, USA); Martti Lehto (National Defence University, Finland); Tara Leweling (Naval Postgraduate School, Pacific Grove, USA); Paul Lewis (technology strategy board, UK); Peeter Lorents (CCD COE, Tallinn, Estonia); James Malcolm (University of Hertfordshire, UK); Hossein Malekinezhad (Islamic Azad University,Naragh Branch, Iran); Mario Marques Freire (University of Beira Interior, Covilhã, Portugal); Ioannis Mavridis (University of Macedonia, Greece); Rob McCusker (Teeside University, Middlesborough, UK); Jean-Pierre Molton Michel (Ministry of Agriculture, Haiti); Durgesh Mishra (Acropolis Institute of Technology and Research, India); Dr. Yonathan Mizrachi (University of Haifa, Israel, Israel); Edmundo Monteiro (University of Coimbra, Portugal); Evangelos Moustakas (Middlesex University, London, UK); Dr. Kara Nance (University of Alaska Fairbanks, USA); Muhammad Naveed (IQRA University Peshawar, Pakistan, Pakistan); Mzukisi Njotini (University of South Africa, South Africa); Rain Ottis (Cooperative Cyber Defence Centre of Excellence, Estonia); Tim Parsons (Selex Communications, UK); Michael Pilgermann (University of Glamorgan, UK); Engur Pisirici (govermental - independent, Turkey); Dr Bernardi Pranggono (Glasgow Caledonian University, UK); Dr. Muttukrishnan Rajarajan (City University London, UK); Andrea Rigoni (Booz & Company,, USA); Dr. Neil Rowe (US Naval Postgraduate School, Monterey, USA); Raphael Rues (DigiComp Academy, Switzerland); Filipe Sa Soares (University of Minho, Portugal); Dr char sample (Carnegie Mellon University/CERT, USA); Prof. Henrique Santos (University of Minho, Portugal); Prof. Chaudhary Imran Sarwar (Mixed Reality University, Pakistan); Dr. Damien

Sauveron (Mathematics and Computer Sciences, University of Limoges, France); Sameer Saxena (IAHS Academy, Mahindra Special Services Group , India); Prof. Dr. Richard Sethmann (University of Applied Sciences Bremen, Germany); Dr. Yilun Shang (Singapore University of Technology and Design, Singapore); Prof. Paulo Simoes (University of Coimbra, Portugal); Prof. Jill Slay (University of South Australia, Australia); Dr Joseph Spring (University of Hertfordshire, UK); Anna Squicciarini (University of Milano, Italy); Iain Sutherland (Noroff University College, Kristiansand, Norway.); Jonas Svava Iversen (Danish Broadcast Corporation, Denmark); Anna-Maria Talihärm (Tartu University, Estonia); Dr. Selma Tekir (Izmir Institute of Technology, Turkey); Prof. Sérgio Tenreiro de Magalhães (Universidade Católica Portuguesa, Portugal); Prof. Dr. Peter Trommler (Georg Simon Ohm University Nuremberg, Germany); Betrand Ugorji (University of Hertfordshire, UK); Craig Valli (Edith Cowan UniversitY, Australia); Rudi Vansnick (Internet Society, Belgium); Richard Vaughan (General Dynamics UK Ltd, UK); Stilianos Vidalis (Newport Business School, Newport, UK); Dr. Natarajan Vijayarangan (Tata Consultancy Services Ltd, India); Dr Sune von Solms (Council for Scientific and Industrial Research, South Africa); Marja Vuorinen (University of Helsinki, Finland); Prof Mat Warren (Deakin University, Australia, Australia); Dr. Kenneth Webb (Edith Cowan University, Australia); Dr. Santoso Wibowo (Central Queensland University, Australia); Dr. Trish Williams (Edith Cowan University, Australia); Simos Xenitellis (Royal Holloway University, London, UK); Dr Hannan Xiao (University of Hertfordshire, UK); Dr. Omar Zakaria (National Defence University of Malaysia, Malaysia).

Biographies

Conference and Programme Chairs

Prof Dr Gabi Dreo Rodosek studied computer science at the University of Maribor, Slovenia. She obtained her PhD and habilitation degree from the Ludwig-Maximilians University in Munich in 1995 and 2002, respectively. Since 2004 she holds the Chair for Communication Systems and Network Security at the Universität der Bundeswehr München. She is Spokesperson of the Research Center CODE (Cyber Defence) and Vice Dean of the Faculty. Since 2014 she is member of the Supervisory and Advisory Board of Giesecke & Devrient GmbH and the Governing Board of the German Research Network (DFN). She is furthermore member of the Executive Committee of the EU No E Project FLAMINGO, member of the Editorial Advisory Board of the International Journal of Network Management, and of several other national and international advisory councils and committees.

Dr Robert Koch received his PhD in 2011 and is now a senior research assistant and lecturer for Computer Science at the Universität der Bundeswehr München (UniBwM). His main areas of research are network and system security with focus on intrusion and extrusion detection in encrypted networks, security of COTS products, security visualization and the application of artificial intelligence. He has several years of experience in the operation of high security networks and systems. His research papers were published in various international conferences and journals. Additionally he serves as program chair and member of the technical program chair for numerous conferences.

Keynote Speakers

Kai Horten is Chairman of the Board of Management (CEO) at ESG Elektroniksystem- und Logistik-GmbH. Before taking on this position, he was CEO of Premium AEROTEC, one of the world's leading companies for the development and manufacture of civilian and military plane structures, and from 2006 to 2011 managing director of ATLAS Elektronik, an innovative, globally active system provider for marine electronics. Mr Horten completed a degree in aerospace engineering (Dipl.-Ing. Univ.) at the Universität der Bundeswehr in Munich.

Dr Detlef Houdeau is the Senior Director for Business Development in the government Identification (ID) market segment under the Chip Card & Security ICs (Integrated Circuits) business group at Infineon Technologies AG. Houdeau first joined Siemens AG in 1985 as the corporate research and development in Berlin, Germany. Over a period of 25 years, he was posted to various sites covering Berlin, Erlangen, Munich and Regensburg in Germany. In 1999, Infineon Technologies AG was spun off from Siemens AG. Born in Germany, Houdeau holds a Diploma in Machine Construction, focusing on micro-system technology in Germany from the Technical University in Berlin, Germany. Promotion at the Technical University in Berlin, Germany, in 1994. He has published over 100 papers and filed 40 patents in the smart card application segment with focus on the government ID market.

Mini-Track Chairs

Dr Olga Angelopoulou, BSc, MSc, PhD is a senior lecturer in computer science at the University of Hertfordshire in the UK. Olga received her PhD from the University of Glamorgan on the 'Analysis of digital evidence in identity theft investigations'. She was previously a senior lecturer at the University of Derby, where she was managing both an undergraduate and a postgraduate programme in digital forensics. Her research interests are in the areas of digital forensics, identity theft and online fraud and information security. She has a number of publications and presentations in the field and is involved in numerous scholarly activities.

Dr. Mils Hills is Associate Professor of Risk, Resilience and Corporate Security, Northampton Business School, UK. Mils specialises in the development of new concepts as a source of options for defence and commercial solutions. He has a background in Information Warfare that began in 1998 (as the first security anthropologist in the UK), by way of heading the UK capability in the human sciences of Information Warfare through secondment to the UK Cabinet Office, consultancy across government and the private sector, research and teaching in university to ongoing support to UK forces.

Dr. Aki-Mauri Huhtinen is a Military Professor at Finnish National Defence University, Department of Leadership and Military Pedagogy, Helsinki, Finland. His expertise areas are military leadership, and philosophy of war. He has published peer-reviewed journal articles, a book chapter and books especially about information warfare and also non-kinetic influence in battle space.

Dr. Martti Lehto, is Adjunct professor, Col (ret.) Martti Lehto has over 35 years of experience as developer and leader of C4ISR Systems in Finnish Defence Forces. He is now a Cyber security and Cyber defence researcher and teacher at the University of Jyväskylä in the Department of Mathematical Information Technology. He also coordinates the Cyber Security MSc. and Doctoral programmes. He has over 60 publications, research reports and articles on areas of C4ISR systems, cyber security and defence, leadership and management, information warfare and defence policy. Since 2001 he has been the Editor-in-Chief of the Military Magazine**.**

Dr. Char Sample is a researcher focused on Threat Intelligence at MITRE and a visiting fellow at the University of Warwick. Dr. Sample has most recently focused her studies on the role of culture in cyber attack and defence behaviours. Additionally, she has interest in metrics, traffic analysis, risk management and measurement, and predictive models. Dr. Sample's background encompasses commercial, government and most recently academic environments. She continues to try to merge the best features of all three environments.

Masterclass Facilitator

Dr Edwin "Leigh" Armistead is the Director of Business Development for Goldbelt Hawk LLC, the Programme Chair for the International Conference of Information Warfare and an Adjunct Lecturer for Edith Cowen University in Perth, Australia. He has written nine books, 18 journal articles, presented 17 academic papers and served as a Chairman for 16 professional and academic conferences. Formerly a Master Faculty at the Joint Forces Staff College, Leigh received his PhD from Edith Cowan University with an emphasis on Information Operations. He also serves as a Co-Editor for the *Journal of International Warfare*, and the Editorial Review Board for European Conference on Information Warfare.

Contributing Authors

Muhammed Bello Abdulazeez is a PhD Student in Computer Science at the University of Liverpool UK. My research is in the area of Intrusion Detection and Prevention Systems in the Cloud environment. My main area of interest is Dynamic protocol Analyses of protocols in the network. Other areas of interest are; Distributed Systems Security, Secure Payment Systems and Telecommunications.

Sakari Ahvenainen is a PhD candidate of the Finnish National Defence University and lieutenant colonel (G.S) (ret.) of the Finnish Army. He has published six international papers on Information Warfare and six articles in the year book "Tiede ja Ase" (Science and Weapon) of the Finnish Society of Military Sciences.

Sultan S Alshameani bachelor degree in computer science from Taif University he has worked for Taif University since 2008. He got his master from the University of Sydney in computer networks in 2012. He is now a third year PhD student at University of Liverpool. His interests are in the Application Isolation, Application Architecture, Virtual Network monitoring and Cloud computing.

Henry Au holds a B.S. and M.S. in Electrical Engineering. He has been with SPAWAR Systems Center Pacific since 2009 and supports the Navy through basic and applied research in the fields of ultra low power electronics, real time image processing, practical applications for fully homomorphic encryption, and cyber security software tool development.

Igor Bernik, Ph.D is the head of the Information Security Department at the Faculty of Criminal Justice and Security, University of Maribor, Slovenia. His research fields are information systems, information security, and the growing requirements for information security awareness in cyberspace.

Ian Bryant is a Professional Engineer, currently Principal Fellow at the University of Warwick, and Technical Director of UK's Trustworthy Software Initiative (TSI). He also contributes to various Standards Development Organisations, including Chairing BSI Panel IST/033/4/4 (an ISO JTC1 SC27 WG4 shadow), and is a Rapporteur for ETSI's MTS Security.

Jim Chen is Professor of Cybersecurity in the Information Resource Management College at the U.S. National Defense University (NDU). His expertise is in cybersecurity technology, cyber strategy, and cyber warfare. He has published widely on these topics. He is a recognized cybersecurity expert.

Allen Church is a practicing cyber security architect and strategist, who has advised a number of U.S. government agencies in this area. He was the first to produce an enterprise security architecture (2003-4) for a U.S. government entity. As the former Deputy Technical Director for INTERPOL's U.S. National Central Bureau.

Paul Crocker has a PhD in Mathematics from the University of Leeds, UK. He is currently at the Reliable and Secure Computation Group of the Computer Science Department at the University of Beira Interior, Portugal and member of the Portuguese Institute of Telecommunications. His research interests are Information Security, Operating Systems and Parallel and Distributed Programming.

Tiago Cruz is Assistant Professor at the Department of Informatics Engineering of the University of Coimbra (UC) since December 2013, also being a senior researcher at the Centre for Informatics and Systems of the UC. His research interests cover areas such as management of communications infrastructures and services, critical infrastructure security, and network function virtualization.

Alan Dinerman LTC is a United States Army officer with a vast experience in the cyber and information environment. His experiences include positions in operational planning and strategic policy development. He holds a BS from West Point and a MS from the Missouri University of Science and Technology. He is currently assigned to the U.S. National Defense University's Information Resource Management College.

Petrus Duvenage is a counterintelligence specialist with extensive practical experience in various aspects of this field. In the course of his career, he served as an officer in the South African Defense Force, the National Intelligence Service and the State Security Agency. Duvenage holds a PhD from the University of Pretoria

Martin Dvorak Is auditor of information security according to ISO 27001 has worked as an IT consultant at KPMG, where he participated in security projects for major companies in the telecommunications industry and the public sector. Besides his job duties he is devoted to postgraduate studies at the Faculty of Informatics and Statistics at the University of Economics.

Dagney Ellison has studied both at the University of Aberdeen, UK and the University of Pretoria, South Africa. Dagney completed her undergraduate degree in Computer Science at the University of Aberdeen in 2013 and is currently working on her Masters in digital forensics at the University of Pretoria.

Adriana-Cristina Enache received her Bachelor's Degree with honors in Computer Engineering from the Military Technical Academy in 2010.. Since 2013 she has been a PhD student at the Politehnica University of Bucharest. The main objectives of her research include proposing new innovative techniques based on computational intelligence algorithms in order to detect and countervail security threats.

Courtney Falk is working on his doctorate of philosophy degree in information security at Purdue University. Between degrees he spent eight years working in first the government sector and then in private industry writing secure code. His current research goal is to apply natural language processing to information security problems.

Eric Filiol is the head of (C+V) research lab at ESIEA, France and senior consultant in offensive cybersecurity and intelligence. He spent 22 years in the French Army (Infantry/Marine Corps). He holds an Engineer diploma in Cryptology, a PhD in applied mathematics and computer science and a Habilitation Thesis in Computer Science. He is graduated from NATO in InfoOps. He is the Editor-in-chief of the Journal in Computer Virology. He has been a speaker at international security events including Black Hat, CCC, CanSecWest, PacSec, Hack.lu, Brucon, H2HC…

Richard Gomm is an Investigations Officer within the Garda Siochana Ombudsman Commission, Ireland. He graduated from the Metropolitian Police College in 2001, obtained Sergeant rank in 2004 and received a MSc in Forensic Computing from UCD in 2012. With over 15 years law enforcement experience, Richard has established himself as a technical authority in the areas of network security and data recovery.

Gerhard Hansch (M.Sc.) joined the department Product Protection and Industrial Security at Fraunhofer AISEC in 2015. Before, he worked at the Bavarian IT-Security Cluster e.V. and the Regensburg University of Applied Sciences. Gerhard studied Computer Science (B.Sc. in 2011) and Applied Research (M.Sc. in 2012) at Regensburg University of Applied Sciences.

Hugh Harney has been designing and building secure and resilient architectures for over 30 years. He is best known for his work with group communication and key management. In particular, he has designed, and standardized, peer to peer, distributed security architectures that use group data delivery as a core technology.

Barbara Hauer received the M.Sc. degree in secure information systems (Sichere Informationssysteme) from the University of Applied Sciences Upper Austria, Hagenberg, Austria in 2006. After her master studies she started investigating information security threats. She is currently pursuing the Ph.D. degree in informatics with the Johannes Kepler University of Linz, Austria.

Pierre Conrad Jacobs has a MSc Information Security, Introduction to Information Security (UNISA). ICT Security Experience: 14 years. Current: Senior Security Specialist at CSIR DPSS. Project Manager for the DOD Information Warfare Assistance Programme (IWAP). February 2013 – Current: CSIR DPSS Senior Security Architect with clients being Department of Defence, and Department of Communication.

Margarita Levin Jaitner is a researcher of Information Warfare in the cyberspace at the Swedish Defence University with a particular focus on Russian operations as well as Fellow at the Blavatnik Interdisciplinary Cyber Research Center (ICRC). She has previously conducted research at the Finnish National Defence University as well as at the Yuval Ne'eman Workshop for Security, Science and Technology in Tel Aviv. Margarita holds an MA degree in Societal Risk Management as well as a BA in Political Science.

Ravi Jhawar is a Post-Doctoral researcher in the Interdisciplinary Center for Security, Reliability and Trust at the University of Luxembourg. His research interests include attack modeling and assessment, cyber situational awareness, and adaptive cyber defenses. He received his PhD in Computer Sciences from University of Milan and MSc in Information Security from University College London.

Audun Jøsang works at the University of Oslo where he teaches and conducts research in cyber security. Before moving to Oslo in 2008 he was Associate Professor at QUT and research leader of the Security Unit at DSTC in Brisbane, Australia. He also worked for Alcatel Telecom in Belgium and for Telenor in Norway.He received a Master's degree in Information Security from Royal Holloway College, University of London, and a PhD from NTNU in Norway

Major Harry Kantola joined the Finnish Defence Forces in 1991. Between 1st of June 2014 and 30th of June 2016 he was appointment as Researcher at the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE), Tallinn, Estonia. Currently he is also appointed to the Finnish Defence Command as a Cyber Defence planner in C5 (J6) branch.

Nickson Karie a Computer Science PhD candidate at the University of Pretoria South Africa focusing on Digital Forensics. Karie has presented several research papers at different international conferences and published others in world known scientific journals. His research interests include Digital Forensics, Cyber Security, Cloud Security, Network Security, Intrusion Detection and Computer/Information Security Architecture

Victor Kebande is a PhD researcher at the University of Pretoria in the field of Cloud Forensic Readiness at the department of computer science, University of Pretoria. He is a member of institute of information technology professionals of South Africa (IIPTSA) and an active member of Information and Computer Security Architectures(ICSA) research group. His research interest are in cloud forensics and internet security

Rauno Kuusisto works for the Finnish Defence Forces. He has contributed as an expert, consultant and manager particularly on the areas of information availability in strategic decision-making, strategy thinking, product development, research purchasing, project portfolio management, network enabled management and leadership in innovative environment, systems thinking, as well as modeling comprehensive challenges. He has published about fifty academic papers, edited books and research reports.

Tuija Kuusisto is a Security Manager at Ministry of Finance and an Adjunct Professor at National Defence University in Finland. She has worked for public organizations and global and national ICT providers. Her area of expertise includes digitalization and cyber security strategies and information management for decision-making. She has about 70 scientific publications in international and national journals, conference proceedings and books.

Rebecca Lee is a senior at George Mason University majoring in Global Affairs and double minoring in IT and Japanese Studies. Since August 2015, she has been conducting multidisciplinary research in the area of cyber warfare with analyzing cyber and kinetic attacks. Her research has been presented at several well known research conferences."

Antoine Lemay is a researcher at École Polytechnique de Montréal in the Department of Computer and Software Engineering, specializing in securing ICS and SCADA networks against threats from nation states. He also has work experience as a security analyst and holds a number of professional certifications, including CISSP, GSEC and GCIH.

Andrew Liaropoulos is Assistant Professor in University of Piraeus, Department of International and European Studies, Greece. He is also the Assistant Editor of the Journal of Mediterranean and Balkan Intelligence (JMBI) and member of the Editorial Board of the Journal of Information Warfare (JIW). Dr. Liaropoulos has published widely in the area of international security, intelligence and cyber security.

Markus Maybaum is an IT professional with more than 20 years of professional experience in the field of Software-Engineering and IT-Security. He holds a masters degree in informatics specializing in IT security. Currently Markus is pursuing a PhD in information technology at Fraunhofer FKIE with a focus on technical aspects of arms control in cyberspace.

Teija Norri-Sederholm is a researcher at University of Eastern Finland. She received her PhD in Health and Human Services Informatics at University of Eastern Finland. Her postdoctoral research interest are in information flow, situational awareness, and multi-authority co-operation.

Lubos Omelina graduated in software engineering from the Faculty of informatics and Information Technologies, Slovak University of Technology in Bratislava in 2009. He is currently a researcher at the Department of Electronics and Informatics, Vrije Universiteit Brussel. His research interests include computer vision, biometrics, face and iris recognition and applications in related fields.

Stacey Omeleze is completing her research for a Masters degree in Computer Science, at the University of Pretoria in the field of Digital Forensics application to proactive crime reduction in South Africa. She is working towards a PhD on developing an Internet of Things (IoT) Forensic Investigation Framework. She is an active member of Information and Computer Security Architecture (ICSA) research group and her research interest spans around Information Security, Digital forensics, IoT, Mobile Forensics, Software Architecture, and Digital Forensic Algorithmic.

Florian Otto studied informatics at Coburg University of Applied Sciences and earned a masters degree. Since 2011 he works in the research project "WISENT" developing data mining methods for network monitoring data in the IT-Security domain.

Jarkko Paavola received the Doctoral degree in technology in the field of wireless communications from University of Turku, Finland. He is currently a research team leader and a principal lecturer with Turku University of Applied Sciences, Turku, Finland. His current research interests include information security and privacy, dynamic spectrum sharing, and information security architectures for systems utilizing spectrum sharing

Schalk Peach is a researcher at the Council for Scientific and Industrial Research specialising in cyber ranges, computer network experiments, and network device testing. He is busy with his Master's degree in Computer Science, focusing on network emulation using Linux container technologies.

Sasanka Potluri Bachelors in Electronics and Communication Engineering in India and received Master's degree in Information Technology from Alpen-Adria University Klagenfurt, Austria. Since 2012 working as a Research Assistant in Otto-von-Guericke University Magdeburg for 3 years in Faculty of Computer Science and after that in Electrical Engineering and Information Technology.

Mikko Räkköläinen is a Ph.D. student in the University of Tampere, School of Management, Tampere, Finland. His research interests include security privatization and military outsourcing and bureaucratic politics. He has previously worked in the Finnish Ministry of Foreign Affairs on the Conference on a Weapons of Mass Destruction Free Zone in the Middle East.

Neil Rowe is Professor of Computer Science at the U.S. Naval Postgraduate School where he has been since 1983. He has a Ph.D. in Computer Science from Stanford University (1983). His main research interests are in data mining, digital forensics, modeling of deception, and cyberwarfare.

Tiia Sõmer is an early stage researcher at TUT. Her research focuses on cybercrime and cyber forensics, with interest on serious games in teaching cyber security. Before academic career, she worked for approximately twenty years in the Estonian army, including teaching at staff college and working in diplomatic positions at national, NATO and EU levels.

Abdulghani Suwan is a final PhD year part time student at School of Computer Science and Informatics Faculty of Technology De Montfort University Leicester, He received his MSc from De Montfort University Leicester. Submit two a research papers and it is with great interest and will be attending for ECCWS conference.

Amie Taal is a remarkably talented and highly driven professional offering over 30 years of experience working with computers and over twenty-four years' experience as a digital forensic investigator, IT Security, eDiscovery and Data Analytics Specialist dealing with civil and criminal matter within the public and private sector including the Big 4 and other accounting firms.

Hein Tun Ph.D student of MIET(Russia). The theme of my research is “Analysis of the security of information systems using hybrid model”.The paper analyses a possibility of Hybrid Modelling using for the estimation of information systems ability to prevent the unauthorised penetration.

Solomon Ogbomon Uwagbole is currently a Part-time Ph.D research student at Edinburgh Napier University, Edinburgh, UK looking at bio-inspired SQL injection attack detection and prevention. He received the B.Sc(Honours) degree in Zoology from Delhi University, New Delhi, India in 1995, and M.Sc. in Distributed Computing from Brunel University, UK.

Maarten van Wieren is a Senior Manager, Deloitte Cyber Risk Services, Netherlands. Maarten specializes in modelling complex systems and leads the cyber risk quantification team. Having a MSc in Financial Risk Management and a PhD in Mathematical Physics, he combines the latest cyber risk insights with extensive experience in financial risk, balance sheet optimization and economic models.

Hein Venter received his MSc in Computer Science from the Rand Afrikaans University and is a senior lecturer in the Department of Computer Science at the University of Pretoria. His research interests are in computer and Internet security. He is also a member of the organising committee of the ISSA and the SAICSIT national conferences.

Stilianos Vidalis is a Senior Lecturer at the University of Hertfordshire. He is lecturing in the subjects of information security, digital forensics and cyber operations. His research interests include information security, threat assessment, network security, effective computer defence mechanisms and intrusion detection systems. He holds a PhD in the area of threat assessment.

Murdoch Watney is professor in the Department of Public Law at the University of Johannesburg, South Africa where she teaches criminal law. She worked as a prosecutor and is an admitted advocate of the High Court of South Africa. She contributed to three textbooks and has published extensively nationally and internationally in law journals on the law of criminal procedure, criminal law, law of evidence and cyber law. She has delivered a number of papers at national and international conferences.

Huseyin Yagci was awarded with a bachelor's degree in Computer Engineering at Yaşar University in 2015. He is currently pursuing his M.Sc. Degree in Yaşar University. Since 2015, He is one of the members of Cyber Security Labs in this university. His research interests are Information Warfare, Network Security, Network Forensics, Malware Analysis and Cryptography.

Cagatay Yucel is a Ph.D. candidate and a Research Assistant at Yasar University. Yucel received his undergraduate Computer Engineering degree from Izmir Institute of Technology in 2009 and earned his M.S. degree in Engineering from Yasar University in 2012. His research interests are Information Theory, Information Warfare, Cyber Intelligence, Cyber Espionage, Cryptography and Computer Forensics.

Failure or Denial of Service? A Rethink of the Cloud Recovery Model

Muhammed Bello Abdulazeez, Dariusz Kowalski, Alexei Lisista and Sultan Alshamrani
University of Liverpool, UK
m.abdulazeez@liverpool.ac.uk

Abstract: One of the dominant paradigms of cloud computing is infrastructure as a service (IaaS), which allows organizations to outsource computing equipment and resources such as servers, storage, networking, as well as services such as load balancing and content delivery networks. For vendors offering IaaS, load balancing is a critical aspect and selling point. One component of load balancing is auto-scaling. This feature allows applications to scale up and down dynamically based on load, performance and 'health' of a virtual machine (VM). It used to take years to grow businesses to millions of customers but now this can happen in months or even days, therefore the ability to access a seemingly infinite amount of resources on demand is very appealing to businesses. The entire cloud model relies on dynamic scalability and configurability because it is not practical to manually configure on-demand services. In this paper we reconsider the scaling of services on the cloud, and consider the definition of 'healthy' scaling, a concept vendors do not formally define. We also look at application layer denial of service (DOS) attacks on application servers running compute services. While there have been extensive efforts to defend the cloud against volumetric DOS using network layer defences, detecting and preventing application layer DOS attacks on the cloud is non-trivial due to the size of cloud and the heterogeneity of applications running. We surveyed some of the key cloud providers that offer IaaS such as Amazon Web Services, Windows Azure, Google Compute Engine, Rack Space Open Cloud, and IBM Smart Cloud Enterprise. We specifically analysed their auto-scaling features and looked at the cost implications for customers. We ask the question, does the monitoring feature of these services differentiate between load increase and Application Layer DOS when making the decision to scale up its services VM?

Keywords: IaaS, auto scaling, cloud monitoring, denial of service, DOS

1. Introduction

Cloud computing provides convenient, on-demand, network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services), which can be rapidly provisioned and released with minimal management effort or service provider interactions (Mell & Grance, 2011). From the definition above we can see that the cloud offers the following characteristics:

- **On-demand self-service:** Available when users request without the need for human intervention.
- **Broad network access:** Available through different devices such as mobile, tablet, personal computers, and any other device with network access.
- **Resource pooling:** Resource assignment should not be fixed. There should be a group of resources for several different users that can be used based on need (load of applications), and if users do not need the resources, they should be available to others.
- **Rapid elasticity:** Some resources should be scaled up or down quickly based on the load of application. Rapid Elasticity is one of the properties we will be discussing in this paper.
- **Measured service:** The use of the service should be measured for payment and monitoring purposes.
- **Programmatic Access:** Users should be able to access services using trusted Application Programming Interfaces (API)

Cloud computing is an evolutionary development of existing computing approaches combined with new technologies. The main three main delivery models are software as a service (SaaS), infrastructure as a service (IaaS), and platform as a service (PaaS). The excessive marketing by cloud computing providers and unfair criticism by cloud sceptics has led to polarized perspectives about cloud security. For example, claims that cloud computing is inherently insecure have are as absurd as claims that cloud computing brings no unique security challenges. The key technologies, which the cloud relies upon, are existing network and virtualization technologies. These technologies can guarantee certain levels of security but still have some security issues. Virtualization technologies inherit a lot of properties from grid computing where interoperability, accounting, dynamic scalability, and security are considered to be marginal. However, cloud computing, although it originated from similar resource sharing purposes, gives priority to features such as interoperability, accounting, dynamic scalability, and security (Foster et. al., 2008). In particular, in this study, we will explore how current cloud platforms handle failure. Status monitoring or accounting is an essential component to the smooth

operation of today's cloud data centres, as quick responses to anomalies, failures, or excessive loads have dire business and performance ramifications.

The rest of this paper is organized as follows. We continue this section by discussing related work, application layer DOS and auto-scaling. Section 2 surveys auto-scaling features of major cloud service providers. Section 3 compares the auto-scaling features of the major cloud providers. In Section 4, we present a new security architecture for cloud computing. We assess the new design and compare it with the existing structure in Section 5. Finally, we draw conclusions in Section 6.

1.1 Related work

Several attempts have been made to study how cloud services monitor failure. Cloud architectures are by definition complex; this is due to the huge amount of resources involved and the need to fulfil the SLAs (Service Level Agreements) of different users requiring different services. We describe the attempts made to monitor how cloud handles application-layer attacks in the next section.

(Ibrahim, et al., 2011) Proposed "CloudSec: A Security Monitoring Appliance for Virtual Machines in IaaS Cloud Model", this is an active, transparent, and real-time security monitoring for hosted VMs in the IaaS cloud. CloudSec utilizes virtual machine introspection (VMI) techniques to provide fine-grained inspection of physical memory. They leverage the fact that volatile memory normally leaves imprints of user or kernel rootkits attacks, even in the case of self-hiding malware. CloudSec actively reconstructs and monitors the dynamically changing kernel data structures instances to enable effective detection and prevention for the kernel data rootkits such as dynamic kernel object manipulation and kernel object hooking rootkits. It is evident that this tool is effective in detecting attacks against the operating system but will not be efficient in detecting application layer attacks.

Other papers that looked at similar issues include: (Chonka, et al., 2011) proposed "Cloud security defence to protect cloud computing against HTTP-DOS and XML-DOS attacks". Reference (Vissers, et al., 2014) proposed "DDOS Defence System for Web Services in Cloud Environment". This defence system is placed at the entrance of the network, the cloud system is designed in such a way that it only listens to requests from the defence system. Also, (Chonka & Abawajy, 2012) proposed a solution for detecting and mitigating HX-DOS attacks against cloud web services, it applies two decision theory methods to detect attack traffic and mark attack message, CLASSIE and ADMU. Finally (Wang, et al., 2012) proposed "Exploiting Artificial Immune System to Detect Unknown DOS Attacks in Real Time". All these papers have one thing in common: they try to detect some form of application-layer DDOS attacks but do not address recovery or the issues of integration to the auto-scaling module to improve resilience.

1.2 Application-Layer denial of service attack

One of the threats common to Internet-based technologies is the Denial of Service (DOS) and Distributed Denial of Service (DDOS) attacks. Authors in (Stewart, 2007) described an HTTP DOS based attack where an HTTP flooder started 1500 threads to randomize HTTP requests to a victim's web server. A channel can also be flooded with XML messages that will prevent legitimate users access to the network. Assuming the communication channel is big enough, flooded web servers can induce loss of web service availability. Another form of devastating attack is manipulating the content of the XML message to cause the web server to crash without actually flooding the communication channel with messages (Vissers, et al., 2014).

The subtle form of attack is the focus of this study. Unlike flooding that requires huge resources or, at least, the control of a substantial amount of computing resources. This particular attack is especially successfully carried out on web applications. The challenge with web applications is that traditional network security solutions such as firewalls and intrusion detection and prevention systems do not adequately address this issue; web applications introduce new security risks that cannot be effectively defended at the network level and require application-level defences (Subashini & Kavitha, 2011).

1.3 Auto scaling

The management of computing elasticity is usually an application specific task and involves mapping an application's requirements to the available resources. The process of adapting resources to on-demand requirements is called, scaling (Lorido-Botran, et al., 2012). Under-provisioning of resources hurts performance

which can lead to SLA violations while over-provisioning results in idle resources which lead to incurring unnecessary costs. Logical thought will be to provision for average load or peak load. Average load planning is cost effective but when peak load occurs performance is negatively affected which can lead to disgruntled customers. Planning for peak load ensures performance never suffers, but it is not cost effective. Table 1 summaries the benefits and the drawbacks of using average or peak load when provisioning applications.

Table 1: Manual scaling techniques

Load	Pros	Cons
Average	Less cost	Poor performance during peak periods
Peak	No negative impact on performance	High cost

Thus, it is necessary to come up with a more sophisticated method for efficient and cost effective way to scale an application's resources according to demand. The act of dynamically scaling based on the demand of applications is called auto scaling. Essentially, there are two approaches to auto-scaling:

- **Schedule based mechanisms:** Here, the cyclic pattern of daily, weekly or monthly workload is taken into account and provisioning is done based on the workload. The drawback of this method is that it cannot handle unexpected changes in loads. Many cloud providers offer scheduled based auto scaling mechanisms; this is especially useful to customers who can reliably predict the load of their applications.
- **Rule-based techniques:** In this approach two rules are created to determine when to scale up or down. Each rule is user defined and the condition is based on target variable, for example if the CPU load is greater than 75%. When this happens pre-defined action is triggered, e.g., adding a new VM. This form of auto scaling is classified as reactive because it waits for application load to increase before it reacts. Other techniques are available that try to anticipate future needs which are called predictive or proactive auto-scaling. Major cloud providers mainly use the reactive rule based auto scaling techniques.

2. Auto-Scaling in the real world

In this section, we will discuss the auto scaling features of the major commercial cloud providers.

2.1 Amazon web services (AWS)

Amazon provides its IaaS through its Elastic Compute Cloud (EC2) (Amazon, n.d.). According to Gartner, AWS has ten times more cloud capacity than the other 14 providers combined. This is up from five times the size of its competitors' capacity last year. It is, therefore, imperative to first look at Amazon EC2 auto-scaling features when assessing auto-scaling technologies.

An important feature of rule-based auto scaling in AWS is CloudWatch. CloudWatch monitors the applications that run on AWS in real time and can be used to collect and track various metrics of the applications for users (Amazon, n.d.) CloudWatch only reports the metrics. No further information is given if there is high resource utilization.

Amazon achieves auto scaling through different approaches:

- If users have a certain load, auto scaling can be achieved by scheduling scaling plans based on the known load changes. This uses the schedule based techniques.
- Another way the EC2 achieves auto scaling is by checking when average utilization of the EC2 is high then more instances are added. Similarly, conditions can be set to remove instances when the utilization is low. No mechanism to check the cause of high utilization of resources. This is a form of reactive rule based auto scaling.
- A more sophisticated way to achieve auto scaling in EC2 is through the Elastic Load Balancing. This helps to distribute instances within auto scaling groups (Amazon, n.d.). This uses CloudWatch to send alarms to trigger scaling activities.

2.2 Microsoft Azure

Initially, Azure was a PaaS provided by Microsoft, offering Webrole and WorkerRole for hosting front-end applications and processing of backend workloads. Recently it has allowed users to deploy a Windows image

prepared offline in the cloud. In this approach called VMRole; users can control the entire software stack and can remotely access the VM. This makes the VMRole effectively an IaaS type of service. Moreover, unlike other PaaS cloud server providers, Microsoft has added APIs and enabled remote desktop connections to log into hosting operating systems, which functions more like a virtual machine (George, 2015).

The Azure platform provides manual scaling through the Azure management portal. Users can also scale application running all the types of Virtual Machines, i.e. Webrole, WorkerRole and VMRole. To scale an application running any of the instances above users can add or remove role instance to accommodate the workload. When applications are scaled up and down, there is no creation or deletion of new instances instead a set created previously are turned off and on from an availability zone (George, 2015)

2.3 IBM smart cloud

IBM uses OpenStack to provide auto-scaling features through a feature called Heat. This feature reduces the need to manually provision instance capacities in advance. Users can use Heat resources to detect when a Ceilometer alarm triggers and provision or de-provision a new VM depending on the trigger. These groups of VMs must be under a load balancer which distributes the load among the VMs on the scaling group. IBM Smart Cloud also allows users to manually scale applications based on the workload. Users are also able to scale applications based on predicted schedule (IBM, 2015).

2.4 Rackspace open cloud

Rack Space is another free and open source service. This is done through OpenStack, initially in collaboration with NASA. However, now several other companies have joined the OpenStack Project including IBM mentioned above.

Rackspace also provides auto-scaling based on pre-defined rules on schedule. It does not provide auto-scaling based on dynamic load changes (i.e., it does not provide rule based auto scaling). Another form of scaling that is provided by Rack Space is Webhook (capability-based URL), in this case scaling occurs when a URL is triggered (Abrahms, 2014). This, however, can only be classified as manual scaling.

Rackspace monitors metrics using the Monitoring Agent (Million, 2015). Agents can be installed on the cloud servers that users want to monitor. The checks that are available for users include HTTP, TCP, ping, memory, CPU, load average, file system and network (Support, 2015).

2.5 Google compute engine

Google compute engine uses managed instance groups to offer auto-scaling capabilities that allow users to automatically add or remove instances from a managed instance group based on increases or decreases in load. To create an auto-scaler, users must specify the auto scaling policy the auto-scaler should use to determine when to scale. Users can choose to auto-scale using the following policies:

- Average CPU utilization
- Cloud monitoring metrics
- HTTP load balancing serving capacity, which can be based on either utilization or requests per second.

Users can also scale cloud the resources up or down manually or use schedules if the can predict future changes in load (Google, 2015).

3. Comparison of the auto-scaling and security features

In this section, we discuss the auto scaling features provided by cloud providers concerning security.

Firstly, all the providers allow users to manually scale their applications based on needs. The second common feature of all cloud providers is schedule based auto scaling where users can predict load increases/decreases at different times, and scaling is done based on these changes. It is a form of proactive auto scaling. In all but one (Rackspace) of the cloud vendors, users can automatically scale their applications based on some Metrics, i.e. true auto-scaling where the resources are scaled on the fly based on current application needs. This, however, is reactive, and it does not predict when the load changes. This can take a while because it takes

between 44.2 to 810.2 seconds to start an instance of a virtual machine, depending on the cloud provider and the type of VM in question (Mao & Humphrey, 2012).

Table 2: Comparison of cloud feature of different manufacturers

Provider/Properties	Manual Scaling	Schedule Based	Rule Based Reactive	Rule Based Proactive	Investigates Reason For High Utilization
AWS EC2	✓	✓	✓	✕	✕
Microsoft Azure	✓	✓	✓	✕	✕
Rackspace	✓	✓	✕	✕	✕
Google Compute Engine	✓	✓	✓	✕	✕
IBM Smart Cloud	✓	✓	✓	✕	✕

Another critical feature the providers are yet to address is the cause of change in load (we are interested in the increase in load in this paper). Where there is an increase in CPU load utilization, HTTP requests per unit time or other metrics, alarms are simply raised by the Monitoring system e.g. CloudWatch in EC2 or Heat in IBM and application capacity is increased. The Monitoring System does not check for the reason of the increased in load it just alerts the auto-scaler to increase resources to accommodate the increase in load. This is a good feature if the event occurred due to increased business activity but it is not fair especially to the cloud customer if it occurs as a result of malicious activity. Figure 1 illustrates the current cloud architecture.

An interesting finding of this survey is in the documentation of EC2 Auto Scaling Developer Guide (Amazon, n.d.). Amazon provides health checks for all its VMs. Virtual Machine in EC2 has two states, Healthy and Unhealthy. However, Amazon does not provide a definition for the states in their documentation. Moreover, there is no mechanism to detect what causes the Unhealthy state of the VM.

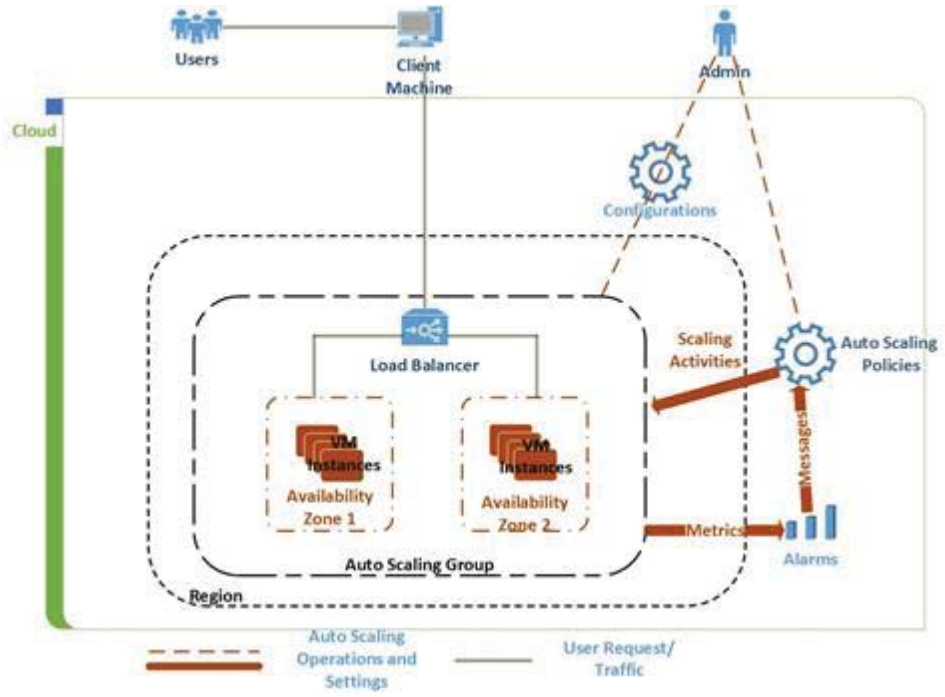


Figure 1: Current cloud architecture

4. Proposed new architecture

The Cloud Vendors have made a lot of effort to monitor the performance of the cloud, to automate all processes of scaling VMs and to provide customers with uninterrupted service. What they have not done yet is to provide a reason for the failure of Virtual Machines. This can be unfair to cloud customers if the reason of a failure is as a result of Application Layer DOS and not the true increase in a load of application. When this occurs, and

resources are scaled up, customers can be overcharged, and occurrences of similar events might not be prevented. In this paper, we will propose two different approaches to tackling this problem.

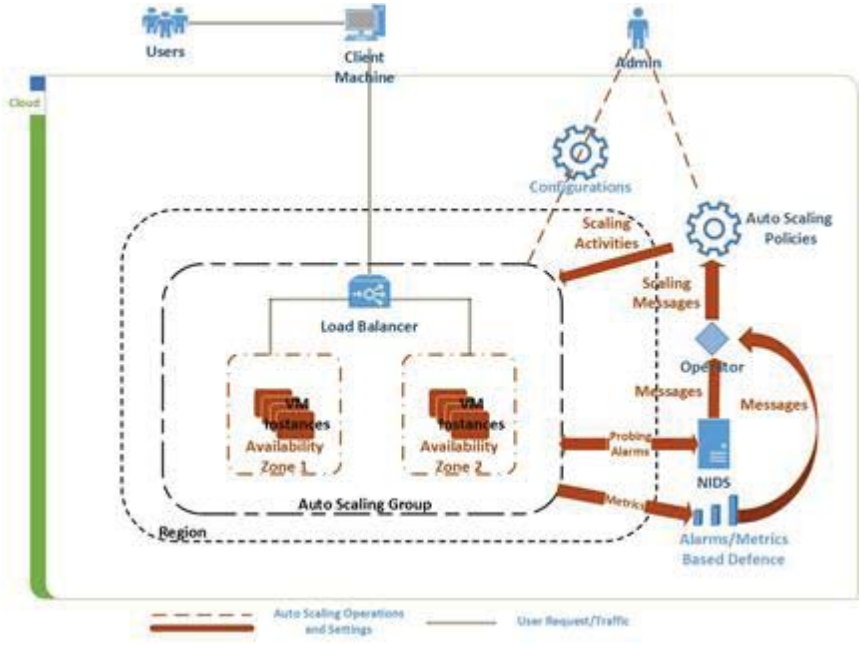


Figure 2: Proposed new architecture

4.1 Use of current metrics

All the cloud vendors surveyed in this paper provide metrics on the use of resources in the cloud. These are important because the metrics can be analysed to determine why a virtual machine failed or why it using a lot of resources. Using two metrics in this paper: average CPU utilization and HTTP load balancing serving capacity, which can be based on either utilization or requests per second.

Let n be the number of HTTP request in a VM time unit,

u Average CPU utilization during that time,

UF be Utilization Factor is UF = n / u

The Utilization Factor will be used to make a decision on whether there is a legitimate increase in load due to more HTTP requests or the growth in load is due to other reasons. Given equal u, higher UF means several request are being severed by VM, which is potentially a good thing, but low UF means few HTTP requests are occupying too many resources which signify potential Application layer attack. This is assuming the volumetric based defences have done their work, to avoid false negative in case of DDOS. Different applications will have different Utilization Factor based on the requirements of the applications. The administrator has the role of assigning the Utilization Factor based on history and application needs.

Other Metrics can be used to achieve similar results.

4.2 Host based intrusion detection system (HIDS)

An alternative way, albeit an expensive one, is the use an HIDS that will analyse every packet coming into the VM to check the possibility of application layer attacks. Features to look at to detect include number of requests, HTTP header inspection, content-length, number of elements, nesting depth, longest element and namespace in a SOAP Message. The possible attacks that will be looked at include Flooding, Header Outlier, Size Outlier, Feature outline and Coercive Parsing. Figure 2 shows the location of both metrics and the HIDS based approaches in the cloud architecture.

5. Assessment

This section provides the comparison between the current and the proposed architectures based on the type of attack that occurs and the speed of auto scaling decision. To have a better comparison we need to define a threshold for normal behaviour depending on application needs, to do that we need to have some sort or data sets that we can use. Unfortunately, the dataset is not available. Therefore, we made the comparison of proposed and existing approaches based on some hypothetical scenarios. The summary of the comparison is that the proposed methods are better alternatives than the current model in all the facets analysed.

5.1 Application layer DOS

Application Layer DOS occurs when a single packet causes exhaustion of server resources. This will normally affect a single VMs in the cloud environment.

Current Architecture: System will scale based on set parameters. In this case, if the increase in load is as a result of Application Layer DOS attack, not genuine customer requests, cloud customer will pay for resources that they do not need.

Proposed Architecture: System scale based on set parameters but a customer will be aware of the reason scaling occurred. The provider will also be aware of the reason for scaling, will be able to stop a future attack, and will scale the system down to the appropriate resources and compensate the customer for lost services.

5.2 Volumetric DOS (flooding)

Volumetric DDOS occurs when the resources of servers are increased due increase in malicious network traffic. It is sometimes distributed because the requests are sent by multiple compromised nodes across the internet. This attack has the potential to affect multiple VMs Clusters or even cloud gateways depending on their severity.

Current Architecture: Scaling will occur as defined by the rules. Here also, if the increase in load is as a result of Application Layer DOS attack, not genuine customer requests, cloud customer will pay for resources that they don't need.

Proposed Architecture: Scaling will occur as defined by the rules, but the customer will be aware of the reason scaling occurred. The provider will also be aware of the reason for scaling, may be able to stop an attack, therefore scaling system down to the appropriate load and compensate the customer.

5.3 Speed of auto scaling

This is a performance based case where we check how quickly auto scaling decision can be made when load increase reaches a predefined threshold.

Current Architecture: Fast based on predefined rules because this can be checked in constant time. This is because

Proposed Architecture: The metrics-based system is equally fast based on simple arithmetic of predefined rules and decision can be made in constant time. However, the HIDS based system is slower because entire packets have to be check. This is linear to the size of the message.

5.4 Resource over-utilization

Increase in load (resource over-utilization) can be a genuine increase in business activity, Application Layer DOS or Volumetric DDOS.

Current Architecture: Resources are simply increased without taking into account why there is an increase in load.

Proposed Architecture: Resources are added, and system investigates whether the increase is due to a legitimate increase in load or as a result of application layer attacks.

6. Conclusions

Cloud computing currently supports many information systems, and it will continue to be used. However, it is crucial to ensure that both the customers and the vendors get their fair share of compensation in the presence of increasing hostility from attackers. In this paper, we introduced scaling techniques and the limitations of scaling techniques for current applications with ever-changing workloads. Most current cloud service providers, provide auto-scaling mechanisms that are better suited for today's dynamic environment, but most cloud providers support auto-scaling techniques that are reactive, and they do not investigate the reasons where applications fail or why there is increased load in their VMs. We then proposed a mechanism for checking the cause of failure. This mechanism is based on the currently available metrics that are already provided to customers while another approach is to use HIDS to detect DOS at the nodes (VMs) in the cloud environment. The future work of this research to test the proposed model and have empirical results to reiterate the advantages as discussed in the initial assessment. Another possible direction is to look other combination of metrics because we only introduced one in this study.

References

Abrahms, M. (2014), '"Rackspace Auto Scale overview"', [ONLINE] Available at: http://www.rackspace.com/knowledge_center/article/rackspace-auto-scale-overview. Rackspace, Accessed 9th October, 2015.

(Amazon), '"Amazon Elastic Compute Cloud"', [ONLINE] Available at: http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/ec2-wg.pdf Accessed 9th October, 2015.

(Amazon), '"Amazon CloudWatch Developer Guide API Version 2010-08-01"', [ONLINE] Available at: http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/acw-dg.pdf Accessed 9th October, 2015.

(Amazon), '"Elastic Load Balancing Developer Guide"', Accessed 9th October, [ONLINE] Available at: http://docs.aws.amazon.com/ElasticLoadBalancing/latest/DeveloperGuide/elb-dg.pdf 2015.

Chonka, A. & Abawajy, J. (2012), Detecting and mitigating HX-DoS attacks against cloud web services, *in* 'Network-Based Information Systems (NBiS), 2012 15th International Conference on', pp. 429--434.

Chonka, A.; Xiang, Y.; Zhou, W. & Bonti, A. (2011), 'Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks', *Journal of Network and Computer Applications* 34(4), 1097--1107.

Foster, I., Zhao, Y., Raicu, I. and Lu, S., 2008, November. *Cloud computing and grid computing 360-degree compared*. In Grid Computing Environments Workshop, 2008. GCE'08 (pp. 1-10). Ieee.

George, A. D. (2015), '"How to Auto-scale an Application"', [ONLINE] Available at: https://azure.microsoft.com/en-gb/documentation/articles/cloud-services-how-to-scale/ Microsoft, Accessed 9th October, 2015.

Google (2015), '"Scaling Based on CPU or Load Balancing Serving Capacity"', [ONLINE] Available at: https://cloud.google.com/compute/docs/autoscaler/scaling-cpu-load-balancing Accessed 9th October, 2015.

IBM (2015), '"IBM Cloud Orchestrator Version 2.4.0.2 User's Guide"', [ONLINE] Available at: Accessed 9th October, 2015.

Ibrahim, A. S.; Hamlyn-Harris, J.; Grundy, J. & Almorsy, M. (2011), CloudSec: a security monitoring appliance for Virtual Machines in the IaaS cloud model, *in* 'Network and System Security (NSS), 2011 5th International Conference on', pp. 113--120.

Lorido-Botrán, T.; Miguel-Alonso, J. & Lozano, J. A. (2012), 'Auto-scaling techniques for elastic applications in cloud environments', *Department of Computer Architecture and Technology, University of Basque Country, Tech. Rep. EHU-KAT-IK-09* 12, 2012.

Mao, M. & Humphrey, M. (2012), A performance study on the vm start-up time in the cloud, in 'Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on', pp. 423--430.

Mell, P. & Grance, T. (2011), 'The NIST definition of cloud computing',

Million, S. (2015), '"Install and configure the Cloud Monitoring Agent"', [ONLINE] Available at: http://www.rackspace.com/knowledge_center/article/install-and-configure-the-cloud-monitoring-agent Accessed 9th October, 2015.

Subashini, S. & Kavitha, V. (2011), 'A survey on security issues in service delivery models of cloud computing', *Journal of Network and Computer Applications* 34(1), 1--11.

Support, R. (2015), '"Available checks for Cloud Monitoring"', [ONLINE] Available at: http://www.rackspace.com/knowledge_center/article/install-and-configure-the-cloud-monitoring-agent Accessed 9th October, 2015.

Vissers, T.; Somasundaram, T. S.; Pieters, L.; Govindarajan, K. & Hellinckx, P. (2014), 'DDoS defense system for web services in a cloud environment', *Future Generation Computer Systems* 37, 37--45.

Wang, D.; He, L.; Xue, Y. & Dong, Y. (2012), Exploiting Artificial Immune systems to detect unknown DoS attacks in real-time, *in* 'Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on', pp. 646--650.

Balancing Mobility Algorithm for Monitoring Virtual Machines in Clouds

Sultan Alshamrani, Dariusz Kowalski, Leszek Gąsieniec and Muhammed Abdulazeez
Department of Computer Science, University of Liverpool, UK
S.alshamrani@liverpool.ac.uk
darek@liverpool.ac.uk
lechu@liverpool.ac.uk
m.abdulazeez@liverpool.ac.uk

Abstract: Amid the rapid growth of Internet users, cybercrime is becoming one of the most challenging tasks for the systems and applications designers to deal with. The cybercrime threat is reflected in the increased number of cases and methods used by criminals. Systems based on cloud computing are natural targets due to their complexity (greater room for security weaknesses) and increasing popularity. Cloud computing is a modern technology that enables users to share resources in a virtual storage and computing environment. A cloud system is based on multiple physical servers. It provides a universal environment with a (large) number of Virtual Machines (VMs) that is available to many users accessing this system via the Internet. This form of access makes cloud systems weaker than physical networks. In order to prevent or minimize the number of attacks and in turn to secure data storage, any malicious behaviour such as external undesirable interventions should be rapidly identified and halted if possible. In this paper, we focus on the discovery of malicious behaviour via determining unwanted symptoms rather than via targeting particular malicious behaviours of the system directly. The main contribution of this paper consists in several new mechanisms for monitoring Virtual Machines and further experimental work targeting efficient ways of visiting VMs in order to discover malicious symptoms. We want to find the fastest and the best set of weights for visiting VMs.

Keywords: virtual machines, malicious behaviour, cloud monitoring, network patrolling, Cloud computing

1. Introduction

Internet users, are often confused with respect to the exact definition of cybercrime. Popular definitions include online crime or computer-related crime (Gordon and Ford, 2006). However, the form of illegal activity changes over time. Cybercriminals modify their methods continuously (Provos et al. 2009). The impact of cybercrime also changes including its costs. In the UK alone payment card fraud over the Internet is estimated at $210 million, and this includes only losses from the banks. These losses have caused 14% of the UK customers to avoid purchases over the Internet (Anderson et al. 2012). Similar numbers refer to the cloud systems' owners and users.

The cloud computing concept usually refers to two aspects. The first one comprises applications and services themselves over the Internet. The second refers to the hardware and software components in data centers offering the respective services (Armbrust et al. 2010). In turn, the purpose of cloud systems is to improve performance of data centers by exposing their virtual services. Those services can be accessed from anywhere in the world and on demand upon the user's requirements (Quality of Service) (Buyya et al. 2009).

Indeed, as a cloud user, there is a definite need to secure personal and workload data. This is even more essential, for those who have valuable and vulnerable data, such as private companies, banks or units from the government sector. Therefore, cloud data storage should be secured and safe. For this reason, it is of the utmost importance to detect any malicious behavior in a cloud computing system. In order to support the efficient discovery of malicious behavior, our research attempts to recognize the symptoms using network-patrolling algorithms, c.f., (Harrison et al. 2012).

This paper starts with a summary of background research. First, the mobility algorithm that is proposed by Harrison et al (2012) and Reduce-Max algorithm that is proposed by our paper in (Alshamrani et al. 2015) will be explained. The main part focuses on the explanation of the work conducted by the authors on monitoring Virtual Machines with the target of discovering malicious symptoms through a number of experiments. We will balance the mobility algorithm by setting two weights in front of each part of the mobility formula. These two weights have a sum of one. The work of the Reduce-Max algorithm is when the first weight is zero. Then we will focus on a comparison of distributed randomized algorithms that patrol the cloud system independently. In this work, there are three sets of weights for the Reduce-Max algorithm. These two are Uniform distribution and

Random distribution. In this paper, we use an additional set that is Poisson distribution. The interesting point is that Poisson distribution gives the best results as shown in this paper.

2. Summary of background research

Cloud monitoring helps to scale resources and utilize the adaptation of the cloud (Shao et al. 2010). In order to monitor symptoms, our approach attempts to use monitoring processes in a distributed fashion reflecting the distributed nature of cloud systems. As a result, a considerable amount of basic computing of cloud systems does not subdivide nicely into small tasks (Birman, K. 2012). Here in our approach, the task of identifying symptoms is subdivided among Forensic Virtual Machines (FVMs). FVMs are tiny virtual machines (VMs) that monitor other VMs in order to identify symptoms. FVMs' job is to recognize the first indication of a threat that could be a virus or other type of malware or misbehavior (Harrison, k. 2010). A symptom is an abstraction of some characteristics that can be related to malicious behavior (Harrison et al. 2012). Symptoms monitoring in this approach is done by using a simple and repetitive mobility algorithm.

Harrison et al (2012) propose the mobility algorithm, which is the most relevant tool for discovering symptoms. To be more specific, the authors propose a class of algorithms that, depending on their parameterization, could be more or less efficient in cloud monitoring. They also pose a question that parameterization leads to efficient monitoring – we address this question in our work.

In the following section, we provide examples of symptoms that direct us to detect cybercriminals' actions and stop them.

2.1 Examples of symptoms

- Modifying the time attributes of a file. One of the malicious behaviours is changing file's attributes (Shabtai et al. 2009).
- Service mix attributes; it could be possible to determine a standard set of inbound or outbound services provided to a particular user. For instance, while he or she is on a business travel, he or she is predicted to use email or file transfer options only. Any other actions on his account, via Telnet from different network's ports, may declare an intrusion (Kazienko, P. and Dorosz, P. 2003).
- Identifying suspicious fragments of code: use of crypto algorithms is very popular with malware writers. Snippets of program code that has been obfuscated or the system containing known crypto algorithms can be a sign of malicious behaviour (Harrison et al. 2012).

3. Model description

Firstly, let us describe the meaning of the following elements: FVMs mean Forensic Virtual Machines and VMs are Virtual Machines. K in this document refers to the number of configurations, N references the number of Virtual Machines (VMs), and M is the number of Forensic Virtual Machines (FVMs). We assume in this paper that we have 1024 VMs. We chose this particular number of VMs to suit the needs of a provider of cloud services as Toosi et al (2011) state that one provider can host 1024 VMs at the same time. All VMs are pairwise connected.

3.1 Description of mobility algorithm

Forensic Virtual Machines (FVMs) choose the best possible VM to inspect. Each FVM carries a copy of an algorithm (e.g., Mobility algorithm) to know the next target Virtual Machine. It is crucial to make it distributed to suit the nature of cloud systems. Harrison et al (2012) state "it is not possible to deploy many FVMs for each VM." They declare this because of the vast scale of the cloud system (Ullah, and Khan, 2014) with limited resources available for security aspects (Wang et al. 2013). This does not contrast with the fact of virtually unlimited computational resources of cloud computing. A cloud service-provider offers numerous resources for cloud customers and users (Xie and Liu, . 2014). However, these resources are for users and should not be wasted on cloud security. Therefore ideally, FVMs should use small resources and avoid costly coordination and centralized computing.

Mobility algorithms deployed in FVMs are looking for a symptom and change their target Virtual Machines repeatedly. Each FVM should look for only one symptom, and all VMs must eventually be visited. The urgency of a VM changes according to how many symptoms are detected in the last visit and the importance of frequent

visits to this VM. Harrison et al (2012) also propose the following formula to assess the urgency of visiting of Virtual Machine *v*:

$$F(v) = \sum_{i=0}^{K} \frac{Disc(c_i,v)}{size(c_i)}\ val(c_i) + \lambda T$$

where $Disc(c_i,v)$ is the number of symptoms in configuration c_i discovered at VM *v*, $size(c_i)$ is the number of potential symptoms in configuration c_i, $val(c_i)$ is how dangerous the symptoms are, λ is an important set for weighting a VM and *T* is the current time minus the last time when this VM was visited by some FVM. Therefore, the weight of a VM increases according to three aspects. The first factor is the appearance of symptoms at this VM on the last visit. The second one is the time from the last visit to this VM to the time of current visit. The final aspect is weighting sets. The last two are considered as one part regarding the waiting time and the importance of VMs.

3.2 Balacing mobility algorithm

In this paper we propose to balance the two parts of the mobility formula. As explained in the introduction, we set two weights before each part of the mobility formula. These weights total one. If the first weight before symptom part equals zero and the second one before second part equals one, then the Reduce-Max algorithm works.

If we balance the mobility algorithm as above, then we can find the fastest way or the fastest method of using sets of data to visit VMs, regardless of the appearance of symptoms. Next is an explanation about the Reduce-Max algorithm.

3.3 Reduce-Max algorithm

We assume that there is a network of *N* VMs, also called nodes. Each node *i* has its own weight (λ_i). The Reduce-Max algorithm schedules an FVM to visit a node *i* with a maximum value of $\lambda_i * t_i$ where t_i is the time since the last visit of some FVM to node *i*.

In the following sections, a number of sets for the factor λ_i of VMs' weights are considered and analyzed in order to see how the Reduce-Max algorithm behaves.

3.4 Input settings of VMs' weights

In our simulations for balancing mobility algorithm, we consider the following three settings for weights of VMs.

3.4.1 Random distribution

Random input here means that the values of weights are selected randomly and independently from integer values from 1 to 10.

3.4.2 Uniform distribution

The start of "uniform distribution" concept was in Hermann Weyl paper in 1916 (Weyl, H. 1916). Then, a sequence is called to be uniformly distributed if one holds (Athreya et al. 1978). Uniform input in this paper means that all VMs have the same set of weights, for example, all of the VMs' sets of weights are ones. Weights of VMs here reduced only in response to the time since the last visit and the appearance of symptoms because weights are equal for all VMs.

3.4.3 Poisson distribution

The Poisson distribution is introduced by a French mathematician, Simeon Denis Poisson. It is a discrete probability distribution that specifies a probability of a given number of events occurring in a fixed time or space (Haight, F. 1967). Poisson distribution used here in this paper as a thought that it could be the best to handle.

Poisson distribution generate numbers for VMs weights according to the following formula

$F(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ (Yatess, R. D., and Goodman, D. J. 1999), where k in this paper is a given number between 1 and 16 according to the number of FVMs being used.

4. Methodology and results

4.1 Methodology of simulations

Simulations are done for the algorithm Reduce-Max, which reduces the highest value of VMs' weights by the fact that some FVM visits this VM. The method of reduction used in this paper is randomized local reduction. We choose this way of reduction rather than deterministic coordinated reduction that is used with randomized local reduction in our paper (Alshamrani et al. 2015) for the following reasons. It is better to use randomized local reduction of specific highest range because it saves time due to the fact that we choose the next VMs locally, from just a range of highest values and do not spend resources on coordination. This is much more efficient than finding, by the deterministic algorithm, the highest weight of a VM from the whole of VMs' weights, as the latter requires coordination mechanisms between FVMs. In addition, recall that the advantage of the former is its lightweight implementation, i.e. not relying on substantial resources or coordination and communication.

In the randomized local reduction, each FVM chooses randomly from a given range of highest weights. The considered ranges of highest values are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16. In each execution of the program, we consider 1 to 16 FVMs. These FVMs reduce highest weights of 1024 VMs. The best result of randomized local reduction, for each number of FVMs is selected among the algorithms using different highest ranges, from 2 to 16.

In the resulting graphs below, cf., Fig. 1, 2, 3, the fifteen columns of different coloured boxes display maximum-weight of Reduce-Max in which FVMs choose VMs using random local reduction from the range of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16 of highest weight VMs, respectively. Sixteen rows correspond to different numbers of FVMs – from 1 to 16. The red boxes – one for each considered number of FVMs – denote the boxes with the smallest maximum weight among the results for random selection from 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16 highest latency VMs, i.e. indicate the best highest range (i.e., column id) to be used by the random local reduction for each considered number of FVMs (row).

4.2 Results

In this section, we give and discuss the results of simulations for each set of input weights.

4.2.1 Random distribution

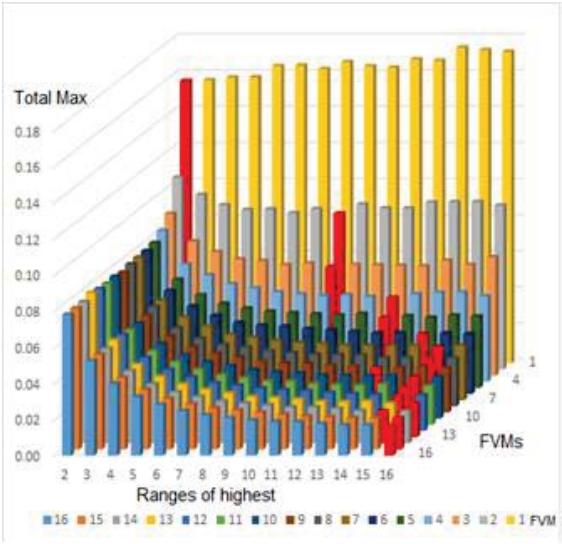


Figure 1: The total max results for randomized local reduction version of Reduce-Max algorithm, taken for fifteen settings of the highest range, applied to random weights

Results in Fig. 1 show that the more FVMs we apply for randomized local reduction, the wider range of highest values achieves the best max. This conclusion comes from the observation that the red box for one FVM indicates the range of two highest values as the best for randomized local reduction; however, the red box moves towards the range of 11 highest values for 2 FVMs and it is in the range of 16 highest values for 12, 14 and 15 FVMs.

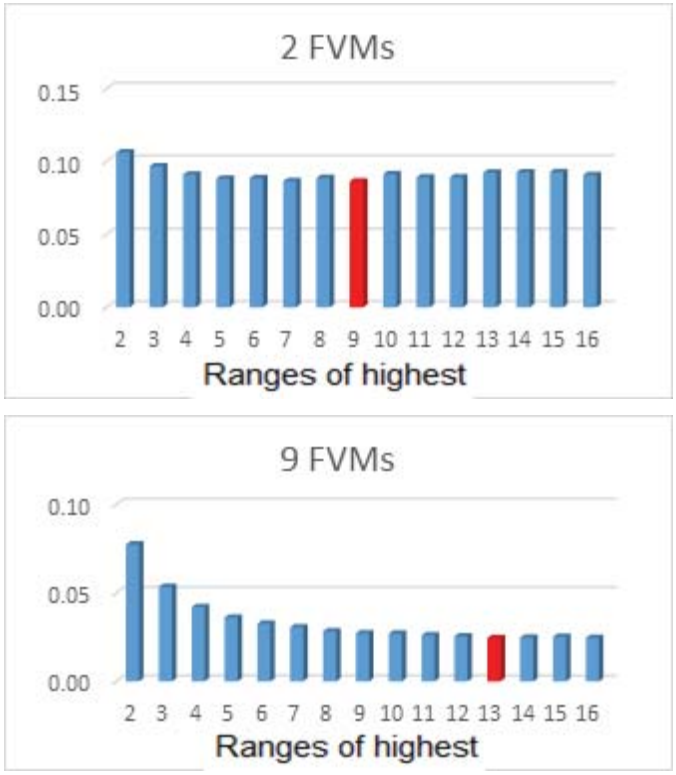


Figure 2: Samples of the total max results for 2 and 9 FVMs taken for fifteen settings of the highest range applied to random weights

Additionally, from Fig. 1, for only one FVM we get worse max weight for a wider range of highest values. It is around 0.16 for the window of two highest and just under 0.18 for the window of 16 highest values. However, these results change with the increase of the number of FVMs. For example, the max weight is about 0.08 for the range of two highest values if there are 16 FVMs, then it goes down to just above 0.02 for the range of 16 highest values.

From the above, the highest total max for random distribution is around 0.18. The red boxes reach the range of 16 highest if only 14 FVMs are applied in the system.

In fig, 2, two slides are taking from fig. 1 to see a deep view into the graph. In this fig. 2, it can be seen clearly from that, the highest window can be reached, when 2 FVMs applied in the system, is 9 highest and 13 highest when 9 FVMs applied.

4.2.2 Uniform distribution

Unlike random weights, for uniform weights the red box, showing the best range for choosing a random VM from, changes from the range of 2 highest through the range of 9 and 13 highest until it reaches 16 highest window for 11, 13, 14, 15 and 16 FVMs.

The Uniform distribution weights, as in fig. 3 below, show better results than the Random distribution in regards to two aspects. The first one is about the number of red boxes in the range of 16. The number of them is 8 whereas they are 3 for random sets. The second one is about the total max itself. The highest total max for uniform sets is just under 0.16. However, it is around 0.18 in the situation of random numbers.

The two slides in fig. 4 are taking from fig. 3 with the same numbers of FVMs to compare among all three sets of weights and to have a deep look inside the graphs. Here for Uniform set, it can be clarified that the highest

window can be reached, when 2 FVMs applied in the system, is 8 highest that is one highest less than the number of highest for Random set. In addition, 15 highest reached when 9 FVMs applied that mean two highest more for Uniform weights.

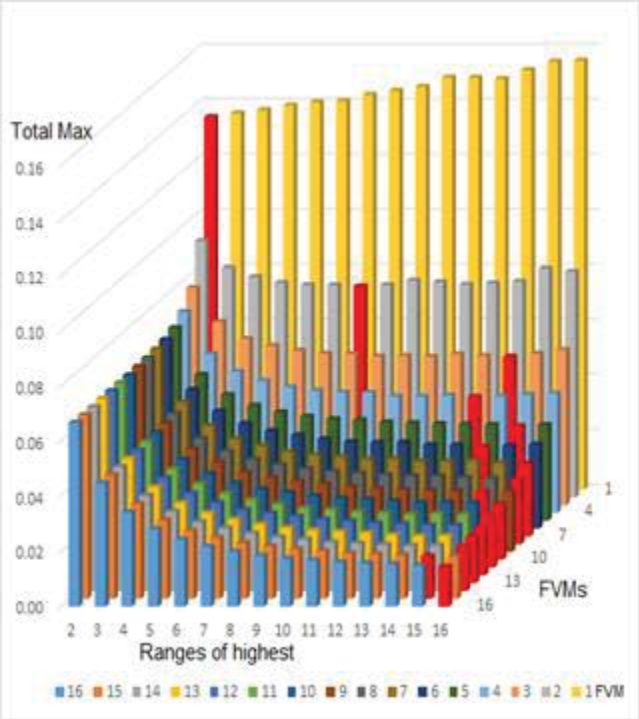


Figure 3: The total max results for randomized local reduction version of Reduce-Max algorithm, taken for fifteen settings of the highest range, applied to uniform weights

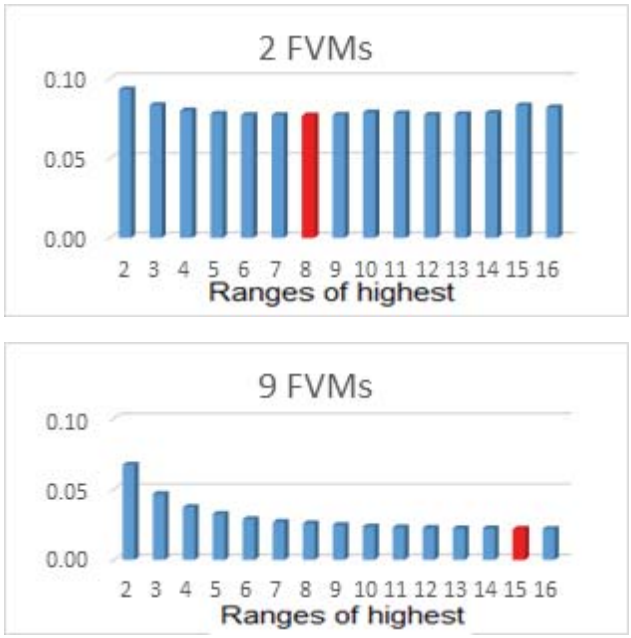


Figure 4: Samples of the total max results for 2 and 9 FVMs taken for fifteen settings of the highest range applied to uniform weights

4.2.3 Poisson distribution

The pattern of red boxes in choosing the best window of highest values for random local reduction for Poisson distribution, c.f., Fig. 5, is different from what was shown previously. It is the best pattern of the three sets considered in this paper. There are 12 red boxes out of 16 in the top last two ranges with 15 and 16 highest ranges. Nine of them are in the highest range in the system.

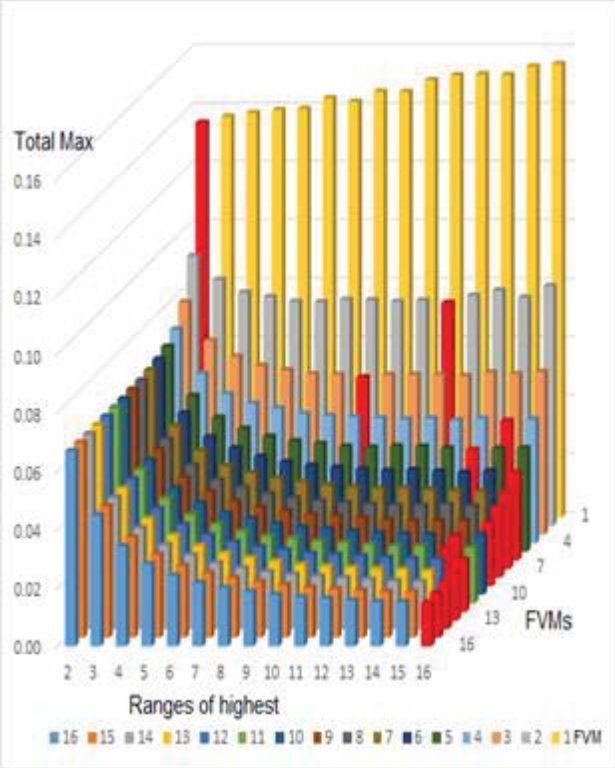


Figure 5: The total max results for randomized local reduction version of Reduce-Max algorithm, taken for fifteen settings of the highest range, applied to Poisson weights

The red boxes are moving from the range of 2 highest through the range of 11 and towards the range of 15 and 16 highest for 1, 2, 3, and 4 towards 16 FVMs. The highest total max for Poisson distribution is the smallest, which is another reason for making Poisson distribution a good choice for most systems. However, it is not far from the one for uniform sets.

The two slides in Fig. 6 are taking from Fig. 5 with the same numbers of FVMs, 2 and 6 FVMs. For Poisson distribution, the highest window can be reached, when 2 FVMs applied in the system, is 12 highest that is better than both Random and Uniform, it was in 9 highest for random and 8 highest for Uniform. Additionally, when 9 FVMs applied for Poisson distribution, it is in the 16 highest window that is the biggest window the system and it is also better than previous sets for the same number of FVMs.

5. In conclusions

In this paper, we attempted to quantify the belief that any malicious behavior of cybercriminals in the cloud could be detected early. We focus on a simplified version of the mobility algorithm, called the Reduce-Max algorithm, and the version using random local reductions, which discover the symptoms of malicious behavior. Three configurations of distribution of weights were considered, implemented and analyzed for the Reduce-Max algorithm.

Overall, the best results for randomized local reduction, which is the version of Reduce-Max requiring little resources and no coordination, occur in the case of Poisson weights. However, the result for the Uniform set of weights is close to the one for the Poisson distribution. The result for random weights is less effective for most cases of applying FVMs than the other two considered distribution.

The future trends could include investigation for more balancing weights of mobility algorithms such as 0.25 for the symptoms part of it and 0.75 for the second part. Additionally, more concerns about first part or "symptoms part" should be considered, for instance, grouping symptoms according to the healing process.

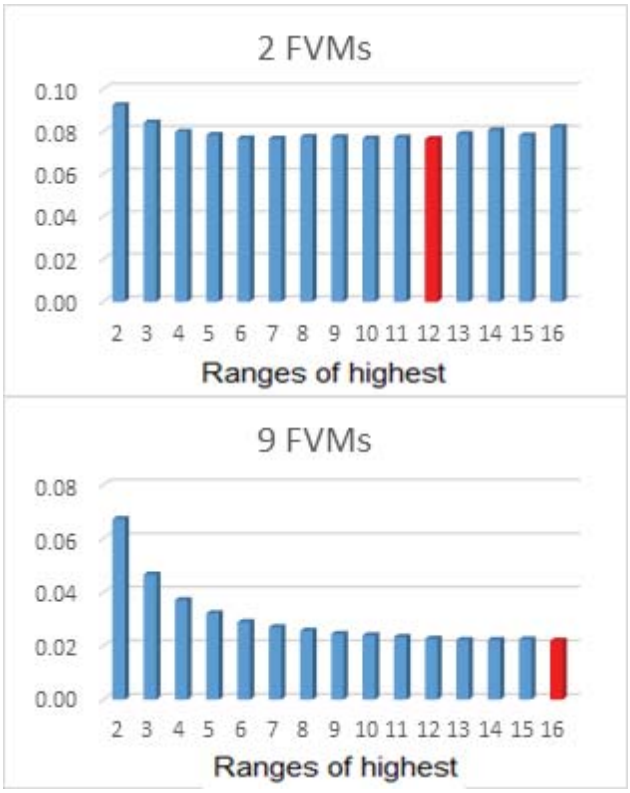


Figure 6: Samples of the total max results for 2 and 9 FVMs taken for fifteen settings of the highest range applied to Poisson weights

References

Alshamrani, S. Kowalski, D. and Gasieniec, L. (2015) "Efficient discovery of malicious symptoms in clouds via monitoring virtual machines," in Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on, pp. 1703–1710.

Anderson, R. Barton, C. Böhme, R. Clayton, R. Eeten, M. Levi, M. Moore, T. and Savage, S. (2012) Measuring the cost of cybercrime.

Armbrust, M. Fox, A. Griffith, R. Joseph, A D. Katz, R. Konwinski, A. Lee, G. Patterson, D. Rabkin, A. Stoica, I. (2010) A view of cloud computing. *Communications of the ACM,* 53(4):50–58.

Athreya, K., McDonald, D. and Ney, P. (1978) "Coupling and the renewal theorem." American Mathematical Monthly :809-814.

Birman, K. (2012) Guide to Reliable Distributed Systems: Building High-Assurance Applications and Cloud-Hosted Services. Springer-Verlag New York Incorporated.

Buyya, R. Ranjan, R and Calheiros, R. (2009) Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. In High Performance Computing & Simulation, 2009. HPCS'09. International Conference on, pages 1–11. IEEE.

Gordon, S. and Ford, R.(2006) On the definition and classification of cybercrime. *Journal in Computer Virology*, 2(1):13–20.

Haight, F. (1967). Handbook of the Poisson Distribution. New York: John Wiley & Sons.

Harrison, K. (2010) Virtual machines, September 30 2010. US Patent App. 13/822,239.

Harrison, K. Bordbar, B. Ali, S. Dalton, C. and Norman, A. (2012) A framework for detecting malware in cloud by identifying symptoms. In Enterprise Distributed Object Computing Conference (EDOC), 2012 IEEE 16th International, pages 164–172. IEEE.

Kazienko, P. and Dorosz, P. (2003) Intrusion detection systems (ids) part i-(network intrusions; attack symptoms; ids tasks; and ids architecture). Retrieved April, 20:2009.

Provos, N. Abu Rajab, M. and Mavrommatis, P. (2009) Cybercrime 2.0: when the cloud turns dark. Communications of the ACM,52(4):42–47.

Shao, J., Wei, H., Wang, Q., & Mei, H. (2010, July). A runtime model based monitoring approach for cloud. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on (pp. 313-320). IEEE.

Shabtai, A. Moskovitch, R. Elovici, Y and Glezer, C. (2009) Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. Information Security Technical Report, 14(1):16–29.

Toosi, A.N. Calheiros, R.N. Thulasiram, R.K. and Buyya, R., (2011, Sept) "Resource Provisioning Policies to Increase IaaS Provider's Profit in a Federated Cloud Environment," High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on , vol., no., pp.279,287, doi: 10.1109/HPCC.2011.44.

Ullah, K. amd Khan, M. (2014) Security and Privacy Issues in Cloud Computing Environment: A Survey Paper. International Journal Of Grid & Distributed Computing [serial online]. April 2014;7(2):89-98. Available from: Computers & Applied Sciences Complete, Ipswich, MA. Accessed December 20, 2014.

Wang, C. Ren, K. Wang, J. and Wang, Q. (2013, june) Harnessing the Cloud for Securely Outsourcing Large-Scale Systems of Linear Equations. IEEE Transactions On Parallel & Distributed Systems [serial online]. June 2013;24(6):1172-1181. Available from: Business Source Complete, Ipswich, MA. Accessed December 20, 2014.

Weyl, H. (1916) Uber die Gleichverteilung von Zahlen modulo Eins. Math. Ann.77, 313-352.

Xie, F., and Liu, F. (2014). Dynamic Effective Resource Allocation Based on Cloud Computing Learning Model. Journal of Networks, 9(11), 3092-3097.

Yatess, R. D., and Goodman, D. J. (1999) "Probability and Stochastic Processes. A Friendly introduction for Electrical and Computer Engineering.".

Multi-Stage Analysis of Intrusion Detection Logs for Quick Impact Assessment

Henry Au, Mamadou Diallo and Krislin Lee
Department of Navy, SPAWAR System Center Pacific, USA
henry.au@navy.mil
mamadou.diallo@navy.mil
krislin.lee@navy.mil

Abstract: As more systems are networked together enterprise security practices focus more and more on securing networks and end devices, each of which pose a potential risk for viruses, external attackers, and exploitations. As well, with the growing reliance on networked devices and systems for daily operations, it is natural that identifying and responding to cyber attacks is a critical piece in enterprise level operations. In order to minimize operational impact, cyber security is typically tackled from two directions. Shrinking cyber gaps and increasing breach preparedness. However, typical defense technologies such as Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), and Security Information & Event Management (SIEM) systems create 10's of thousands of alarms in minutes, thus creating a new problem. With so many events being logged, detected threats to critical systems can easily be hidden and lost. As well, mining through hundreds of thousands of security logs require long lead times and analysts with extensive knowledge of monitored systems, network topology, and networking protocols in order to accurately identify cyber threats. In this paper we propose, a new cyber analyst tool, Cyber Impact Assessment (CyberIA), to facilitate the processing of IDS logs to quickly determine the impact due to cyber events. The CyberIA system is a data-driven intrusion detection log analysis tool capable of processing tens of thousands of alarms. The system is comprised of two phases. The first phase clusters alarms using the k-means algorithm on a Graphics Processing Unit (GPU) for improved performance. The second phase consists of a supervised machine learning support vector machine (SVM) algorithm which is used to reduce the number of logs presented to analysts. Timing analysis of the k-means clustering algorithm will be conducted on data from DARPA's 1999 intrusion detection dataset. The sequential k-means algorithm will be compared with the GPU counterpart to assess the practical speed up of the first phase algorithm in the CyberIA framework. The beauty of the CyberIA system is that it is a web based framework allowing for quick integration to other data sources and data processing libraries. With the performance increase, the final CyberIA system will process IDS logs, organize them based on impact to security posture, display the information, and ultimately allow analysts to quickly interpret groups of events rather than individual alerts.

Keywords: cyber security, intrusion detection system, clustering, machine learning, log minimization, GPGPU

1. Introduction

Current solutions rely on a combination of IDS, IPS, and SIEM technologies to identify cyber threats to network systems based on various physical and virtual network sensors. These traditional cyber security solutions often generate large sets of log data which can hide true threats being detected. As networks become larger and larger, the generated alarms become a "Big Data" storage and access problem. Not only is access time an issue, but prioritizing which threats should be analyzed first becomes a challenge. To add to that, it becomes increasingly more difficult to correlate previous persistent threats. With what seems like favorable odds to those who seek to harm network systems, it becomes of utmost importance to utilize the data being collected by IDS such as Snort®. Presented in Figure 1 is an example of an IDS alarm generated by Snort®.

```
(Event)
        sensor id: 0    event id: 1     event second: 920898013 event microsecond: 194424
        sig id: 8       gen id: 125     revision: 1     classification: 3
        priority: 2     ip source: 206.48.44.18 ip destination: 172.16.112.100
        src port: 1054  dest port: 21   protocol: 6     impact_flag: 0  blocked: 0

Packet
        sensor id: 0    event id: 1     event second: 920898013
        packet second: 920898013        packet microsecond: 194424
        linktype: 1     packet_length: 80
[    0] 00 10 7B 38 46 32 00 C0 4F A3 58 23 08 00 45 00  ..{8F2..O.X#..E.
[   16] 00 42 9D 00 40 00 7F 06 47 FE CE 30 2C 12 AC 10  .B..@...G..0,...
[   32] 70 64 04 1E 00 15 00 17 AD 57 00 17 AF 17 50 18  pd.......W....P.
[   48] 21 A2 21 89 00 00 70 6F 72 74 20 31 39 39 2C 31  !.!...port 199,1
[   64] 39 39 2C 31 39 39 2C 31 39 39 2C 30 2C 38 30 0A  99,199,199,0,80.
```

Figure 1: Typical Snort® intrusion detection system alert

With facilitating analysis in mind, the CyberIA tool was developed in order to help cluster and prioritize IDS alarms for cyber analysts. As many other systems and algorithms focus on reducing false positives, the aim instead is to create a tool which allows cyber analysts to quickly and easily look at similar alerts in a more intuitive nature. Thus, relying on humans for interpretation and analysis and software for organization and presentation.

2. CyberIA system

The CyberIA framework was developed in order to quickly and easily implement highly optimized C/C++ and GPU based algorithms for data processing. It is a client server web based tool which uses the open source CppCMS development framework. This framework was selected due to it's capability in handling extremely high loads. The C++ nature of the framework also allows for a much simpler integration of Graphics Processing Unit (GPU) based algorithms. All that is required is a simple addition of the library header file and path to library during the compilation process. There is no need for bindings such as those required for Java and NVIDIA® Compute Unified Device Architecture (CUDA™). This section will describe the major components utilized in the CyberIA system.

2.1 Client server web architecture

The CyberIA system architecture is built on a typical client server model. The client is able to access the service via a web page and the server handles the user requests. In this initial version of the system the start and stop time (time window) is specified by the user. IDS logs within the time window are retrieved from a database, clustered using the k-means algorithm, and finally graphed. Again, it should be noted the CyberIA system is meant to facilitate the discovery of cyber intrusions by cyber analysts and is a tool in cyber defense. Figure 2 below depicts the typical Client Server Web architecture employed.

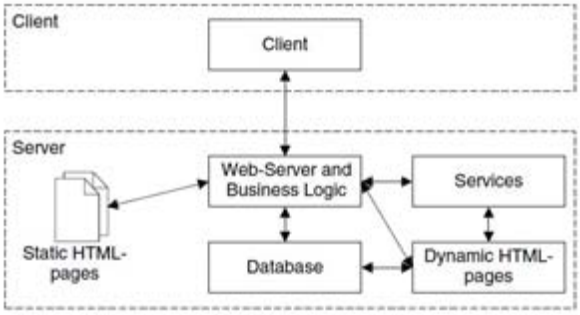


Figure 2: Typical client server web architecture

2.2 Snort® intrusion detection system (IDS)

The database portion of the CyberIA system is populated from the alerts generated by Snort®. Snort® is an open source network intrusion prevention and detection system capable of performing real-time traffic analysis and packet logging on Internet Protocol (IP) networks. It is also able to operate in various modes for purposes of reading network packets, logging network packets, and more typically used network intrusion detection. While operating in the intrusion detection mode, the program will analyze traffic against user defined rules. The power of Snort® is that it uses simple descriptions to create rules which can be easily tailored for a specific network. There are community rules available as well.

As with any IDS system, network traffic needs to be present in order for alerts to be generated. DARPA's freely available 1999 network intrusion data set with labeled attacks is used. Snort® version 2.9.7.0 is used to process the network traffic. Using Barnyard2, an open source interpreter for Snort® binary output files, the alarms were saved into the commonly used open source MySQL® database. Table 1 details the Unified2 data fields and data size for each alert.

Table 1: Snort® Unified2 IDS event structure

Unified2 IDS Event	
Snort Field	Size
sensor id	4 bytes
event id	4 bytes
Snort Field	Size
event second	4 bytes
event microsecond	4 bytes
signature id	4 bytes
generator id	4 bytes
signature revision	4 bytes

Unified2 IDS Event	
classification id	4 bytes
priority id	4 bytes
ip source	4 bytes
ip destination	4 bytes
source port/icmp type	2 bytes
dest. port/icmp code	2 bytes
protocol	1 byte
impact flag	1 byte
impact	1 byte
blocked	1 byte
mpls label	4 bytes
vlan id	2 bytes
padding	2 bytes

As well as logging each alert triggered in Snort® the packet which caused the alert is also logged and describe in Table 2.

Table 2: Snort® Unified2 event packet structure

Unified2 IDS Event Packet	
Snort Field	**Size**
sensor id	4 bytes
event id	4 bytes
event seconds	4 bytes
event microseconds	4 bytes
link type	4 bytes
packet length	4 bytes
packet data	<variable length>

2.3 MySQL® database

MySQL® is the selected database used in the CyberIA framework to store the Snort® generated alerts. The wide use in many web applications requiring relational database allows for ease of integration to other datasets for future CyberIA development. With other aggregated network data the CyberIA system would facilitate even further the cyber forensics recognizing and responding to cyber attacks.

2.4 General purpose computing on graphics processing units

When dealing with large data sets typical with any IDS system, a scalable and relatively inexpensive Graphics Processing Unit (GPU) can be utilized for powerful performance gains. GPUs were originally developed to perform the parallel mathematical calculations required for rendering images. However, GPUs continued to evolve from simple shape accelerators to performing complex computations such as those required for 3d rendering. Only as recently as 2007 did General Purpose Computing on Graphics Processing Units (GPGPU) become widely used for high performance computing. This is largely due to NVIDIA® CUDA™ and OpenCL providing the necessary back end coordination required for managing the hundreds of parallel cores available on GPUs.

Although there are hundreds of cores available on GPUs, it does take work to properly utilize their power. This typically means conducting timing analysis of traditional sequential algorithms. Sanders et al (2011) describes the added complexity due to the overhead associated with GPU initialization and memory bandwidth, the transfer rate between the host CPU and device GPU. Before conducting timing analysis, a simple rule of thumb is to determine whether the problem is task parallel or data parallel. Task parallel refers to problems in which threads are able to work on different tasks independently. Where as data parallel refers to problems in which threads are performing the same computation but on different parts of data. When sequential algorithms fall into the latter, GPUs excel in providing performance increase. As mentioned earlier a limitation of GPUs is their memory bandwidth. Thus, it is important to determine whether problems are bound by memory or computation. Figure 3 depicts the thread, block, grid, and core relationship used in CUDA™. It is easy to see that when hundreds of processing cores are available, data parallel tasks benefit the most.

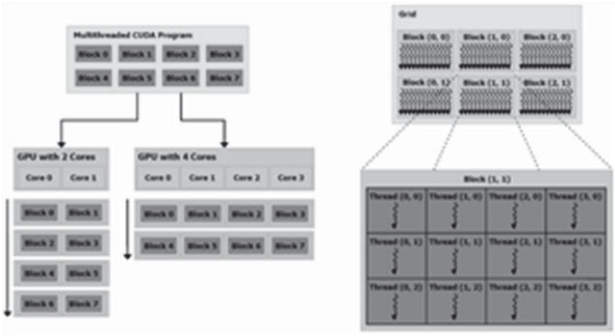


Figure 3: Compute unified device architecture (CUDA™) thread distribution

Memory bound algorithms are those in which the computation done for each data element is minimal and the bottleneck in performance is due to memory transfers. The other, computation bound, refers to problems in which the majority of the processing is occurring due to computations. Algorithms which are data parallel and computation bound make GPUs the ideal choice for increasing performance. It should be noted that this is a general rule of thumb and timing analysis should be conducted to further assess. CUDA C Best Practices Guide recommends profiling applications for hotspots to quickly determine candidates for parallelization. As will be discussed later, the highly data parallel nature of the k-means clustering outweighs the fact that the data being processed is memory bound.

2.5 CyberIA process flow

As described previously, the CyberIA tool processes Snort® IDS alarms, clusters, and binary classifies the alerts in turn reducing the number of alarms presented to the analyst. These alarms are saved in the Unified2 format which is parsed and stored in a MySQL® database. Of the available event fields in the alerts, the time stamp, source IP, and destination IP are of significance for the clustering portion of the CyberIA system. However, for the support vector machine (SVM) machine learning portion, the following features from the Snort® event fields are used, source IP and port, destination IP and port, and signature ID. It should be noted that the focus of this paper is on the CyberIA system and the k-means clustering. The SVM is referenced in order to describe the overall CyberIA data use and flow. Hachmi et al (2013) describes a method to improve intrusion detection based on data mining algorithms in order to reduce the rate of false positives. A similar approach was adopted for the CyberIA system and Figure 4 describes the CyberIA process flow and features of importance for the k-means clustering and SVM. As well, Remya et al (2013) describes a similar process using self-organized feature maps and support vector machines for improved network anomaly detection.

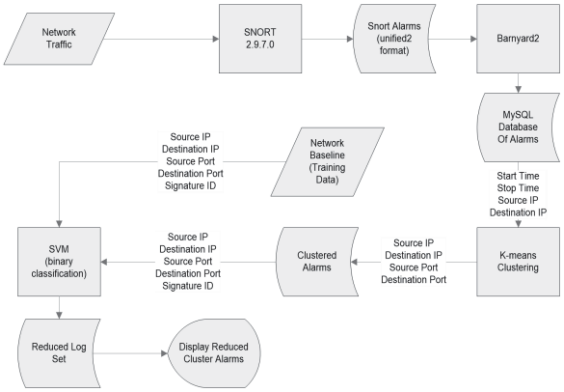


Figure 4: CyberIA data flow process

3. K-Means clustering

In this section the sequential k-means clustering algorithm, feature set, cluster size selection, and parallelized k-means algorithm is discussed. The focus will be on the optimizations performed on the GPU enabled k-means algorithm with detailed timing test results presented in section 4.

3.1 Problem definition

The basic k-means algorithm is comprised of the seeding stage followed by the labelling stage. During the seeding stage, an initial clustering *C* is created by selecting *k* number of points as the initial centroids from the set of data points *N* where C_j is a subset of *N*, j=1,...,k. After the initial clustering stage, the labelling stage is executed where each data point, n_i, an element of *N*, is assigned to the nearest cluster C_j for which the Euclidean distance $D(n_i,c_j)$ is minimal. After assignment each centroid C_j is then recalculated by the mean of all the data points within the particular centroid assignment. This process is repeated until the centroids no longer change.

3.2 Sequential k-means

From the top level description of the k-means clustering algorithm, it is immediately apparent that the Euclidean distance calculation for each node is data parallel. The k-means clustering can also either be memory bound or compute bound depending on the selection of the cluster size. In either event, the sequential algorithm lends it self to exploitation on the GPU. Liao (2013) depicts a detailed implementation of the k-means clustering algorithm which can be seen below in Figure 5.

```
N:  number of data objects
K:  number of clusters

objects[N]:  array of data objects
clusters[K]:  array of cluster centers
membership[N]:  array of object memberships

kmeans_clustering( )
1   while δ/N > threshold
2     δ ← 0
3     for i ← 0 to N-1
4       for j ← 0 to K-1
5         distance ← | objects[i] - clusters[j] |
6         if distance < dmin
7           dmin ← distance
8           n ← j
9       if membership[i] ≠ n
10        δ ← δ + 1
11        membership[i] ← n
12      new_clusters[n] ← new_clusters[n] + objects[i]
13      new_cluster_size[n] ← new_cluster_size[n] + 1
14    for j ← 0 to K-1
15      clusters[j][*] ← new_clusters[j][*] / new_cluster_size[j]
16      new_clusters[j][*] ← 0
17      new_cluster_size[j] ← 0
```

Figure 5: Sequential K-Means algorithm

Although the calculation is not complex, it does fall into the category of being data parallel and computationally bound when solving for large data sets. How large will be discussed later. As with development of any sequential algorithm, a basic understanding of the processing time is desired. Before moving forward with the GPU k-means development, TAU 2.24.1 was used to profile the sequential k-means algorithm using 11,952 IDS alerts, features, source IP, destination IP, and time stamp, and a cluster size of 2,390. It will be discussed later why these particular parameters were selected. The data points were generated using, DARPA's Intrusion Detection Data Sets (online), specifically the March 3rd Transmission Control Protocol (TCP) data dump. The test environment used for the sequential k-means timing is described in Table 3.

Table 3: CyberIA framework test environment

Operating System	Ubuntu 14.04 LTS
Laptop	Lenovo T440p
CPU	Intel Core i7-4600M
Memory	8 GB Memory
GPU	Nvidia GeForce GT 730M
Cores	384
Driver	CUDA Driver 7.0, Runtime 7.0
Version	CUDA Capability 3.5

A few simulations confirmed the sequential k-means taking 68.2 seconds spent 95% of its time computing the Euclidean distance and 5% spent re-computing the centroids. As stated by Shi (1996), one could use Amdahl's (1976) or Gustafson's (1988) Law to determine the theoretical speed up using p number of processors since the "two laws are in fact identical"(Shi 1996). However, taking into account the limitations of the GPU architecture and the overhead associated with memory transfers between host and device, Zechner (2009) was able to

achieve a 14 times speed increase using NVIDIA's G80 GPU family when comparing to a fully SIMD optimized CPU implementation.

3.3 Parallel k-means and optimizations

Having performed a quick timing analysis on the CPU sequential based k-means algorithm the process of moving the Euclidean distance calculation from CPU to GPU should be less complex due to the data parallelism. However, blindly porting over the distance calculation without understanding the GPU nuances will ultimately end in poor performance.

As described in the NVIDIA CUDA Programming Guide (online) there are various optimization techniques when it comes to GPU programming such as reducing unnecessary memory transfers between host and device, using page locked memory for faster access time, utilizing asynchronous memory copy and computations, and creating multiple threads for multiple concurrent memory copy and data computation. Five different versions of the k-means clustering algorithm were developed to compare the processing time for the same data set used for section 3.2.

The "GPU No Opt" refers to the k-means algorithm which parallelizes the Euclidean distance calculation through a bulk calculation, e.g. $D(n_1,c_j)$ for k centroids. The bulk result is copied from GPU back to CPU and the minimum distance determined on the CPU in order to assign n_1 to the nearest centroid. In doing so we incur costly memory allocations, transfers, clean up, and GPU kernel launches each time the kernel is invoked. In our test this would mean at minimum 11,952 memory allocations, memory transfers, kernel executions, and memory de-allocations. Seen in Figure 6 are the costly overhead incurred between each bulk distance calculation.

Figure 6: Non-Optimized K-Means GPU kernel (GPU no opt)

The "GPU Opt" version of the k-means algorithm instead reduces the previous GPU call for each n_i calculation into a single call. In doing so fewer memory allocations, transfers, and kernel executions are required. Instead of creating and destroying memory it is created once and used for each solution of the k-means clustering problem. This dramatically reduces the overhead cost and begins to provide the performance increase desired when using GPUs. Figure 7 depicts a much more efficient kernel Euclidean distance calculation.

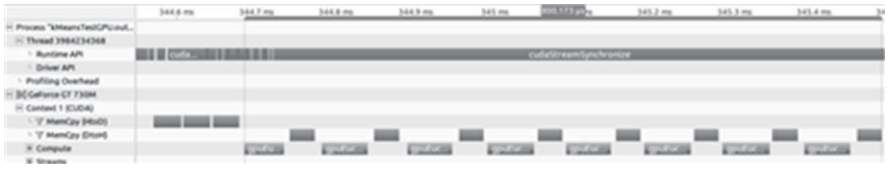


Figure 7: Reduced memory transfer k-means GPU kernel (GPU opt)

Taking the optimized GPU k-means kernel another step further we can begin to "mask" the overhead costs of CUDA memory transfers from device to host. This is done in CUDA using the available concurrent copy and kernel execution and employed using what is known as streams. In essence the bulk result of $D(n_1,c_j)$ for k centroids is copied over while $D(n_2,c_j)$ for k centroids is computed. This in turn reduces the overall processing time. Stream amounts of 2, 8, and 16 were conducted to determine the performance increase. One would immediately think, why not create a many threads as there are data points for peak performance. The simple answer lies in the GPUs limited hardware resources. Using concurrent memory transfer and kernel executions requires hardware resources that may be occupied thereby not guaranteeing concurrent execution; such as when multiprocessors or registers are unavailable. The optimal scenario for this particular test environment where concurrent memory transfer between device and host are masked by the Euclidean Distance kernel execution is presented in Figure 8.

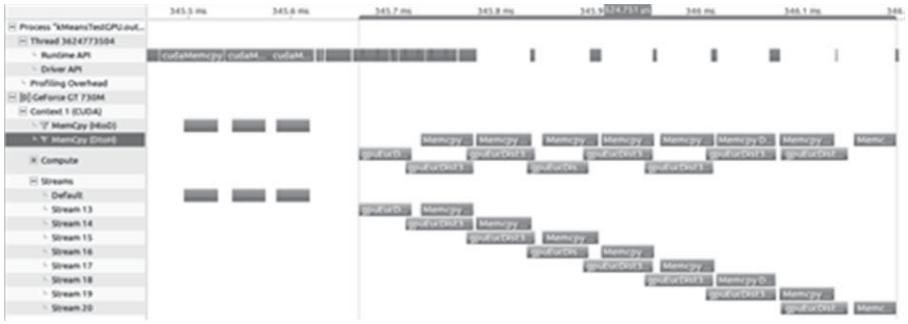


Figure 8: Asynchronous calls with 8 Streams k-means GPU kernel (GPU 8 Strm)

Table 4 describes the different speed up achieved when moving from the CPU based k-means algorithm through the different optimization strategies for GPU algorithm development described previously. For this particular test environment and GPU capabilities, utilizing 8 streams with concurrent memory transfer and kernel executions resulted in a performance gain of 5.3 times. From this point on, the "GPU 8 Strm" k-means implementation is used for the remainder of the timing analysis and performance comparison with the CPU based k-means.

Table 4: CPU & GPU K-Means Optimizations (1,952 IDS alerts, k= 2,390)

Processor	K-Mean(sec)	Euclidean Dist. Calc. (sec)	Centroid Recalc. (sec)	K-Mean Speed Up
CPU	68.15	65.02	3.13	-
GPU No Opt	185.02	181.75	3.27	0.4
GPU Opt	21.9	18.76	3.14	3.1
GPU 2 Strm	15.75	12.6	3.15	4.3
GPU 8 Strm	12.87	9.65	3.22	5.3
GPU 16 Strm	22.06	18.9	3.16	3.1

4. Snort® IDS log clustering

In this section, details regarding the Snort® alerts used for the k-means CPU and GPU implementations will be discussed as well as the performance increase attained through parallel processing using GPUs. As already mentioned, the k-means clustering algorithm solves the partitioning problem of data points through centroid assignment and update using Euclidean distances as a measure of similarity. This allows for each data point to be any dimension *i* for the data set of points n_i in *N*. The number of features, cluster size, and centroid selection are also discussed in this section.

4.1 Feature selection

For the purposes of cyber defense 3 dimensions, or what is referred to as features, are used to describe one IDS alert. More features could be selected to create very specific partitions but for the purposes of the CyberIA system creating partitions that facilitate the discovery of cyber intrusions is the goal. Thus, who and when are the initial primary concern and it follows that the source IP, destination IP, and time stamp are used. These three features provide a good context for who was talking to whom, when, and provides a good basis for partitioning. Take for example port scanning. The act of port scanning itself does not constitute a cyber attack, but it does signify suspicious behaviour, especially when TCP/IP requests from the same source IP using various ports occur within minutes. From this example it is clear that clustering alarms based on these three features provides significant insight into possible intruders. Other factors such as the cluster size and centroid selection also play a role in the end result.

4.2 Cluster size selection

The cluster size *k* selection is an independent problem when dealing with the k-means clustering. The value itself does not prevent the solution of the k-means clustering from being determined but it does affect the end result. Too high a value of *k* and every item essentially becomes it own cluster and too low a value and not enough clusters are formed. For the development of CyberIA the cluster size *k* was based primarily on, *p*, the number of items per cluster we expect to be formed within each partition; *k=Total Points/p*. A range of *k* values was simulated using a large dataset, DARPA's 1999 intrusion week 1 data. The results are discussed in a later section.

4.3 Centroid selection

As well as having control of the features and initial cluster size the centroid selection plays a role in the number of iterations required for centroid convergence. Typically centroids are randomly generated points within the data set N. Since CyberIA is meant to be a tool for analysts, generating random centroids results in different partitions for each clustering request. In an effort to preserve repeatability, the centroid generation is conducted by selecting data points every multiple of k. This oversimplified case of centroid selection works well for IDS alerts, particularly since source and destination IPs are distinct values. Thus, for the remaining timing analysis all initial centroids are generated in this fashion.

4.4 K-Means timing results

Timing analysis on various cluster sizes was conducted in order to determine a realistic performance gain when clustering IDS alerts. Figure 9 depicts the results conducted between the CPU and GPU implementation of the k-means clustering for various cluster size.

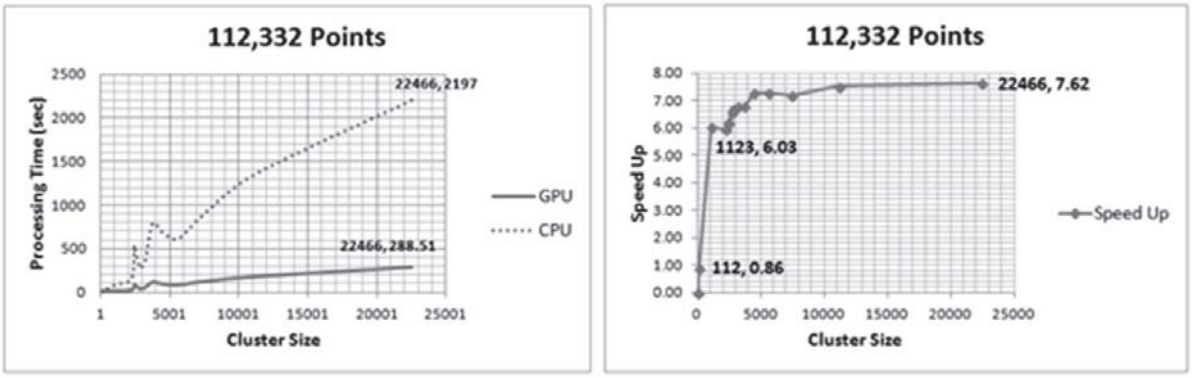


Figure 9: CPU vs GPU processing time for various cluster sizes

As cluster size is varied, processing time can vary between tens of seconds to tens of minutes. Based on CPU and GPU comparisons it was found that for our particular test environment processing 112,332 data points with 112 clusters offered a reduction in performance. This is expected as the memory overhead associated with GPU kernel execution outweighs the data parallelism attained during the Euclidean Distance calculation. It can be seen when dealing with large data sets the GPU k-means excels with as much as a 7 times speed up when *p* is 5. Inspection of the final clusters also proved that the partitioning provided desired source IP to destination IP communications clustered within similar time frames.

Having established *p=5* provided the greatest performance increase while maintaining a manageable cluster size, various datasets were used to compare the CPU k-means to GPU k-means implementation. Table 5 describes in detail the results. It should be noted that all timing conducted includes all GPU overhead with the exception of GPU initialization time. The initialization time is specifically omitted since it is a one time cost which can occur outside the CyberIA System; such as a start-up process during OS boot time.

Table 5: Speed Up of K-Means for various data sets, *p=5*

# IDS Alarms	CPU (sec)	GPU (sec)	Speed Up
8116	22.0	4.8	4.6
11952	68.6	12.4	5.6
12146	23.9	4.3	5.5
12777	26.1	4.5	5.8
13609	63.8	11.8	5.4
14331	68.8	12.0	5.7
21062	157.9	26.1	6.0
21724	231.7	36.1	6.4
22444	104.4	16.5	6.3
27076	94.3	15.5	6.1
28515	178.1	27.5	6.5
29914	193.7	29.5	6.6
31293	468.6	68.9	6.8
35765	201.9	31.0	6.5

# IDS Alarms	CPU (sec)	GPU (sec)	Speed Up
51547	345.9	50.6	6.8
112332	2194.6	288.5	7.6

By exploiting the data parallelism of the Euclidean distance calculation the GPU k-means implementation provides greater performance increase as the data set increases.

5. Operation and evaluation

This section will describe the overall CyberIA framework integrating all components required for the GPU based k-means clustering to occur as a web service as well as system evaluation.

5.1 CyberIA operation

Integrating all components, CppCMS, MySQL®, Snort®, Barnyard2, and the GPU enabled k-means library into the CyberIA framework the CyberIA system was tested. The same test environment used for the k-means clustering analysis was also used for the CyberIA System. The implementation of the clustering portion of the CyberIA system is presented in Figure 10.

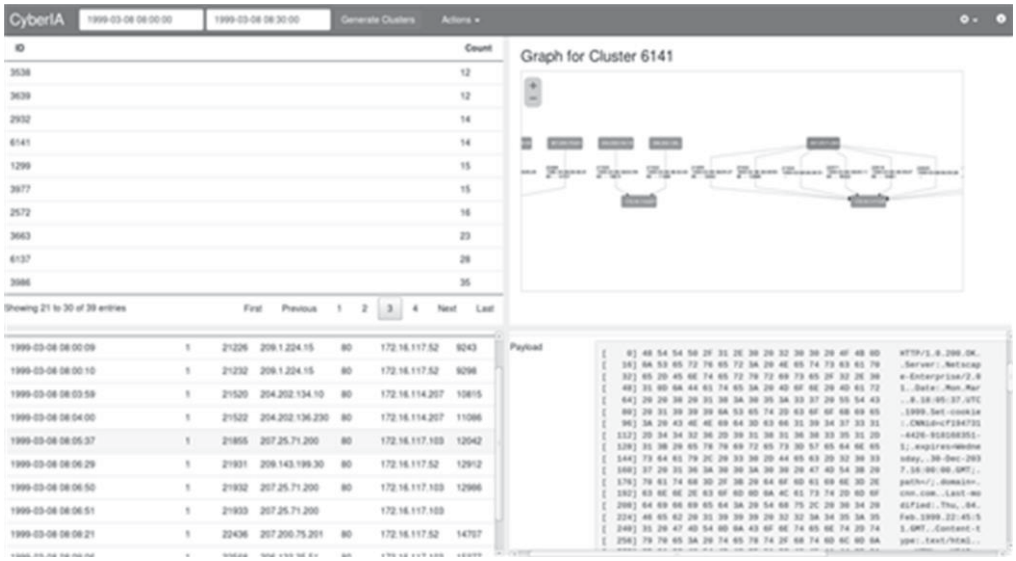


Figure 10: CyberIA K-Means clustering system GUI

The user inputs the desired start and stop time and requests the cluster generation to start. The service then requests data from the database, parses for the source and destination IP and timestamp for the specified time window. After which the data is clustered (labeled) using the GPU enabled k-means. The label is then saved into a new table and stored in the database for this particular cluster generation request. Once complete the clusters are presented to the user and the other fields populated. The other fields present the graph of the clusters, a tabular view of the alerts within the cluster and detailed information regarding each user selected alert. This becomes the starting point for cyber forensics to take place and for human interpretation to begin. Again, it is emphasized that the CyberIA System is meant to be a tool which facilitates the human discovery of possible cyber intrusions.

5.2 Evaluation

The system is able to quickly cluster events and display the results, however the database query and update becomes the bottleneck taking 90% of system time. Since no efforts have been conducted to optimize database queries and inserts it was expected that system performance would be limited by database performance. It should be noted that system performance was previously limited by the CPU based k-means clustering algorithm.

6. Conclusion

Exploiting the GPU for k-means clustering of the IDS alerts proved to be successful with up to 7 times performance increase. Centroid selection and cluster size determination have also been discussed and shown to allow broad selections while maintaining system performance. It is also shown that in order to properly exploit

the parallel processing power of GPUs, optimizations and data structures need to be considered. However, no formal testing has been conducted to analyze how well the k-means algorithm is partitioning data. Things such as inter-cluster and intra-cluster distances have yet to be conducted, though may not particularly provide much added benefit as source and destination IPs vary greatly. Especially in the presence of network address translations. The success of the GPU enabled k-means library with all components proves that a successful web framework has been developed. Future work will involve optimizing database queries, updating graph visualizations for added flexibility, and allowing users with more control over feature selection and cluster size. More investigation into initial centroid generation based on source IPs, destination IPs, and timestamps will be conducted as well.

References

Amdahl, G. (1976) "Validity of the single processor approach to achieving large scale computing capabilities." *AFIPS Spring Joint Computer Conference*.

CUDA C Best Practices Guide, [online], http://docs.nvidia.com/cuda/cuda-c-best-practices-guide/

"DARPA Intrusion Detection Data Sets", [online], https://www.ll.mit.edu/ideval/data/

Gustafson, J. (1988) "Reevaluating Amdahl's Law", *Communications of the ACM*, Vol. 31 No. 5.

Hachmi, F. and Mohamed, L. (2013) "A Two-Stage Technique to Improve Intrusion Detection Systems Based on Data Mining Algorithms", *5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO*, 28–30 April, pp. 1–6.

Liao, W. (2013) "Parallel k-means Data Clustering", [online], *http://users.eecs.northwestern.edu/~wkliao/Kmeans/index.html*

NVIDIA CUDA Programming Guide, [online], http://docs.nvidia.com/cuda/cuda-c-programming-guide/

Remya, R. and Anil, S. (2013) "A Hybrid Method Based on Genetic Algorithm, Self-Organised Feature Map, and Support Vector Machine for Better Network Anomaly Detection", *Fourth International Conference on Computing, Communications and Networking Technologie, ICCCNT,* pp. 1–5.

Sanders, J., and Kandrot, E. (2011) *CUDA By Example, An Introduction to General-Purpose GPU Programming*, Addison-Wesley, Copyright NVIDIA Corporation.

Shi, Y. (1996) "Reevaluating Amdahl's Law and Gustafson's Law", Temple University.

Zechner, M. (2009) "Accelerating k-means on the Graphics Processor via CUDA", *Intensive Applications and Services*, 20-25 April, pp 7-15.

Compliance With Information Security Policies in the Slovene Insurance Sector

Igor Bernik
University of Maribor, Faculty of Criminal Justice and Security, Slovenia
igor.bernik@fvv.uni-mb.si

Abstract: In a globally interconnected cyber space, an appropriate level of information security (IS) is one of the business requirements for the successful operation of business systems, particularly in the fields of high-tech and finance. These two fields are also the most exposed to abuse and different types of cybercrime. The development of information security policies (ISP) is one of the basic measures for providing adequate levels of IS. These must be appropriately promoted within organisations, which ought to provide relevant training courses for employees in order to raise their awareness and make them understand that they are contributing to safe and secure working processes and thus to the general success of the organisation by complying with prescribed procedures and implementing prescribed working methods consistently. The present research study, which focuses on IS and employees' compliance with ISP, aims to determine the state-of-affairs with respect to IS in Slovene insurance undertakings, the types of in-house actions that are or could be implemented to improve the aforementioned issues, the types of measures that are used to decrease the long-term costs of insurance undertakings' operations. It presents analysed data in the field of insurance, measures for improving the level of IS with the best cost-benefit ratio and individual improvements.

Keywords: information security, management, insurance sector, Slovenia

1. Introduction

The business environment is characterised by unpredictable changes, pressures, aggressive rivalry, informatisation and global connectivity, which is why normal functioning depends on information systems. The success of business systems is thus closely linked to the decision-making process in the field of information security management. This is a precondition for the long-term survival and competitiveness, which is also confirmed by research into the development of attacks on information technologies (The impact of cybercrime on business, 2012; Data breach investigation report, 2014; Internet security threat report, 2014) and extremely negative impacts of information incidents on organisations' value (Kshteri, 2006; Bojanc and Jerman-Blažič, 2008; Goel and Shawky, 2009; Son, 2011). This is why researchers focussing on information security emphasise that information security ought to develop as a strong business activity and not as an exclusively technical task (Chang and Ho, 2006; Mishra and Chasalow, 2011; Rhee et al, 2012; Baskerville et al, 2014; Feng et al, 2014).

The viability and security of an entire organisation depends on the quality of (information) security and threat management – information systems and technologies bring about great benefits for an organisation, but they may also become the main source of its shortcomings, if managed and used ineffectively. The reason for this lies in the fact that the combination of numerous information incidents and unfavourable economic conditions creates a paradoxical and complex situation for the management of organisations. Organisations require a high level of information security if they wish to avoid the severe consequences of such incidents. At the same time, they need to rationalise resources dedicated to the management of information security. It is precisely the lack of financial resources that continues to greatly hinder and restrict information security (Global corporate IT security risks, 2013; Sans Institute, 2013; Defending yesterday…, 2014). An overview of current activities in the field of efficient information security provision clearly shows that a lack of such endeavours may also be observed in the scientific research sphere. Apart from market analyses, which may be commercially motivated, there are some comprehensive approaches to the evaluation of organisational practices (e.g. Saleh, 2006; Chang and Ho, 2006; Ponemon Institute, 2010; Sans Institute, 2013), but research studies are generally dispersed, target-oriented and incomparable. They focus primarily on the identification of links between individual phenomena and elements of the system and probability.

Information security is obstructed by the lack of resources, knowledge and competences, as well as indifference. In the scope of the decision-making process, organisations should first evaluate the current state-of-play, identify their security needs and only then proceed with the adoption of decisions and the allocation of resources (Stewart, 2012). In order to contribute to the understanding of research and organisational challenges related to information security management, the present paper presents preliminary findings of a research study that focussed on different factors influencing compliance with in-house rules and regulations or with data protection

rules in different Slovene insurance undertakings. The research provides an insight into individual factors, which have the strongest impact on the compliance with information security policies and consequently into the state-of-play of information security culture.

2. Methodology

Respondents were asked to answer questions related to the implementation of the information security policy and their perception of information security when conducting every-day work in an online survey. Research findings are based on responses provided by 580 employees of different Slovene insurance undertakings, whereby the response rate in individual insurance undertakings ranged from 15 to 50 per cent of employees.

Data were processed by using the PLS (partial least squares) method for structural equation modelling. The model fits the data relatively well (the SRMR value, which represents an indicator of the fitness of the model, amounts to 0.048, which is below the recommended minimum value of 0.08) and the quality criteria have been met, which means that the results are statistically representative and serve as a reliable indicator of the state-of-play. The author is thus able to draw conclusions regarding the general climate and prepare measures for improving individual areas, thus providing for increased awareness and higher levels of information security and organisational culture.

The questions were put together according to individual constructs representing the attitude towards the information security policy and the employees' intention of complying with it when performing day-to-day work. Constructs indicating compliance with the information security policy (ISP) include:

- Attitude to Comply, ATT,
- Benefits of Compliance, BC,
- Costs of Compliance, CC,
- Costs of Non-Compliance, CNC,
- Information Security Awareness, ISPA,
- Intention to Comply, ITC,
- Normative Behaviour, NB,
- Self-Efficiency, SEC.

Constructs and relationships between them were designed on the basis of the theory of planned behaviour developed by Icek Ajzen (Icek, 1991).

3. Situation with respect to information security policy's acceptance

Initially, the author attempted to identify potential relationships between constructs and preliminarily processed results (the discriminant validity test shows that constructs are actually different) and compare mean values (min. 1, max. 7).

Results presented in Table 1 demonstrate relatively high mean values stemming from the attitude towards compliance with the information security policy and employees' intention to comply with it. According to the mean values, the largest "reserves" can be found in improving the benefits of compliance, normative behaviour and the management of non-compliance costs or measures for reducing such costs, as well as in raising information security awareness.

Table 1: Relationships between constructs, mean value of all responses (Source: Own data)

Constructs and variables	ATT	BC	CC	CNC	ISA	ITC	NB	SEC		Mean		Collinearity statistics (Outer VIF)
AT1	0.870									6.219		2.508
AT2	0.835									5.914		2.151
AT3	0.905									5.974		3.078
AT4	0.882									6.157		2.560

Constructs and variables	ATT	BC	CC	CNC	ISA	ITC	NB	SEC		Mean		Collinearity statistics (Outer VIF)
BC1		0.806								3.531		3.698
BC2		0.856								3.790		4.398
BC3		0.925								4.347		3.561
BC4		0.873								4.641		2.685
CC1			0.897							3.724		2.708
CC2			0.799							3.776		1.942
CC3			0.944							3.272		2.478
CNC1				0.921						4.980		4.003
CNC2				0.920						4.861		3.807
CNC3				0.909						5.109		3.316
CNC4				0.830						4.481		2.249
GISA1					0.847					5.417		2.872
GISA2					0.719					4.731		1.935
GISA3					0.804					5.486		2.395
ISPA1					0.857					4.760		4.450
ISPA2					0.875					4.905		5.174
ISPA3					0.896					5.036		4.179
ITC1						0.948				6.540		4.669
ITC2						0.966				6.629		6.875
ITC3						0.946				6.574		4.738
NB1							0.881			5.767		3.200
NB2							0.936			6.152		4.606
NB3							0.871			6.378		3.300
NB4							0.919			6.253		4.012
SEC1								0.967		5.080		7.324
SEC2								0.974		5.034		8.961
SEC 3								0.961		5.173		6.075

The verification of the discriminant validity of constructs according to the Fornell-Larcker criterion (Table 2) confirms the actual differences between constructs.

Table 2: Comparison according to the Fornell-Larcker criterion (Source: Own data)

	Composite Reliability	Fornell-Larcker criterion							
		ATT	BC	CC	CNC	ISA	ITC	NB	SEC
ATT	0.928	0.874							
BC	0.923	0.469	0.866						
CC	0.913	-0.246	-0.085	0.883					
CNC	0.942	0.460	0.579	-0.036	0.896				
ISA	0.932	0.507	0.371	-0.083	0.393	0.835			
ITC	0.967	0.491	0.263	-0.159	0.361	0.438	0.953		
NB	0.946	0.458	0.339	-0.065	0.452	0.360	0.444	0.902	
SEC	0.978	0.417	0.349	-0.044	0.440	0.602	0.353	0.395	0.967

The research also confirms the validity of the model showing relationships between individual constructs: all relationships are statistically significant and expressed in adequate values. Relationships between individual constructs are illustrated in Figure 1.

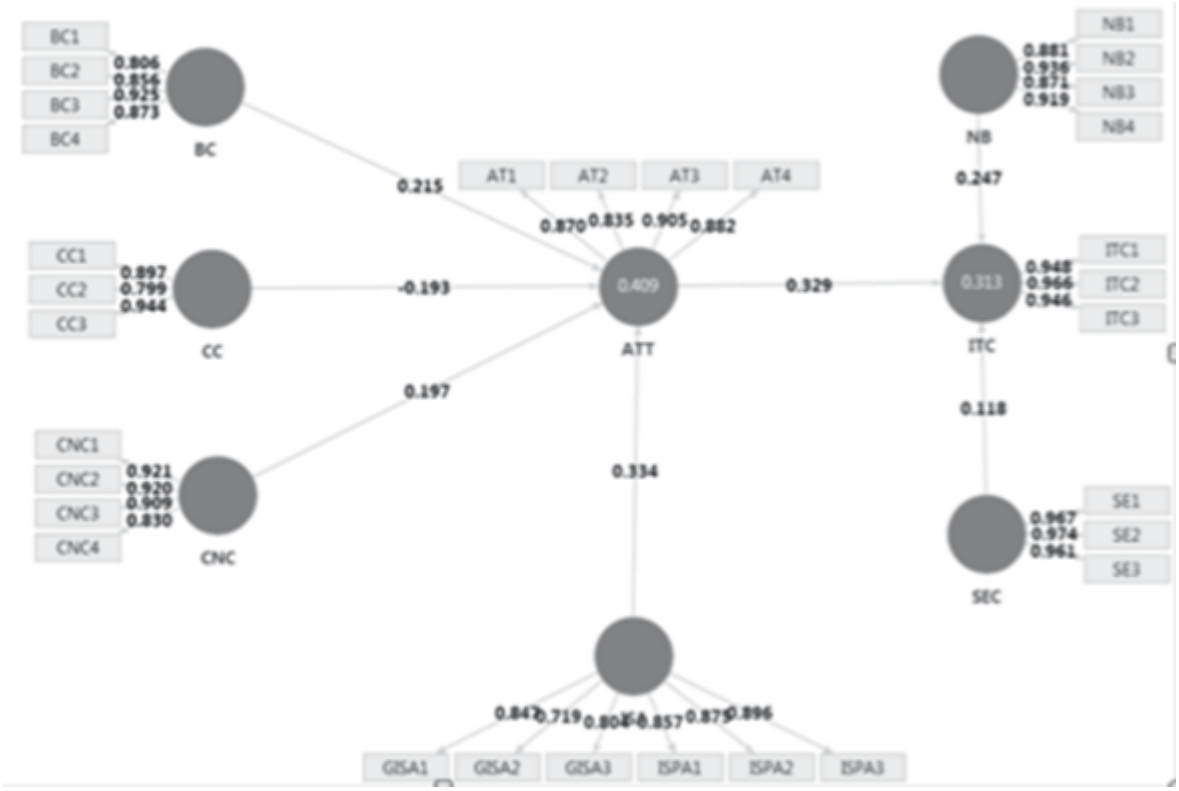


Figure 1: Relationships between constructs (Source: Own data)

The above depiction of relationships between constructs clearly shows that the ATT construct, i.e. the Attitude to Comply with the ISP, represents the most important node, while the ITC construct, i.e. the Intention to Comply with the ISP, is mainly influenced by the Normative Behaviour (NB) and Self-Efficiency (SEC).

Data presented in Figure 1 make it possible to draw conclusions as to the factors that bear the greatest influence on the attitude towards compliance with the information security policy and employees' intention to act in line with it. This is the basis on which the management takes a decision to adopt certain measures and organisational procedures aimed at increasing the level of information security management and thus reducing adverse impacts that cybercrime and cyberwarfare have on the level of threat, which organisations are exposed to due to harmful activities in cyber space, and at decreasing internal negative factors.

The author also analysed potential impacts on the aforementioned constructs that could arise from gender, age and level of education. The author found that the situation matches the basic model of relationships between constructs fully and did not observe any statistical deviations. In other words, gender, age and level of education have no influence on the degree of threat posed to organisations or to the management of information security and employees' compliance with the information security policy.

A comparison between insurance undertakings indicates that individual constructs influence the attitude of employees towards compliance with the information security policy and their intention to act in line with it in a different manner: some focus primarily on benefits generated by compliance with the information security policy, while others prioritise the costs of non-compliance. At the same time, the expectations of important others with respect to compliance with the information security policy influences employees' intention to act in accordance with it, even though this motivation is stronger in some respondents than in others.

4. Conclusion

The above results represent the basis for examining the field of information security and compliance with information security policy. The author finds that the state-of-play in analysed insurance undertakings is good,

while slightly poorer knowledge and awareness, as well as compliance with the information security policy according to individual constructs were identified in some areas. Since the implementation of information security strives to achieve a uniform situation in individual sub-areas and since the analysis of results shows certain deviations, the following paragraphs emphasise some areas that could be improved.

As can be observed from Figure 1, the main area that must be considered further relates to the employees' attitude to comply with information security policy. Since this is a central factor contributing to the provision of a high level of information security within an organisation by its employees, it should be dealt with accordingly. The appropriate management of employees, organisational processes and technologies allows for this factor to be modelled adequately. It is thus important to allocate some financial and time resources towards achieving this aim (e.g. Global corporate IT security risks, 2013; Sans Institute, 2013; Defending yesterday…, 2014). As demonstrated by individual research studies (e.g. Baskerville et al, 2014; Feng et al, 2014; and others), it is reasonable to develop information security as one of the core business functions and not merely as a technical activity. When considered as a business function, information security also contributes to higher levels of awareness among employees and lower financial investments in comparison with those situations where it is considered as a purely technical activity.

The second important construct refers to the intention to comply with the information security policy, which depends on normative behaviour and individuals' self-efficiency. Normative behaviour may be achieved in practice by way of example provided by important others (superiors, colleagues, information security experts), while self-efficiency may be obtained through adequate training and consistent development of human resources.

The issue of compliance with the information security policy shows that employees' understanding of the benefits of compliance (Table 1) has been underestimated or that the information security policy is incomprehensible for users or that users simply feel overly restricted by the prescribed procedures. Employees are often unaware of the costs that may incur as a result of non-compliance, which clearly shows that information security awareness needs to be improved. Untapped potentials were also found in individual areas, particularly in the field of management and increased awareness among employees.

Trends demonstrated by the results of research into the state-of-play in the fields of information security and employees' compliance with information security policies indicate that that the situation in Slovene insurance undertakings is good in both fields. At the same time, results show that there are numerous internal potentials for improving individual areas of information security, particularly for improving certain elements pertaining to the organisational culture and the implementation of information security as a business function. The introduction of proposed measures contributes to a long-term reduction of operational costs in individual insurance undertakings. Further analyses focus on identifying detailed measures at the level of the insurance sector and the presentation of results accompanied by adequate measures for individual entities, which participated in the research, with the aim of improving the general situation regarding information security in this sector.

References

Ajzen, I. (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50 (2): 179–211.

Baskerville, R. et al. (2014) Incident – centred information security: Managing a strategic balance between prevention and response. *Information & Management, 51*(1), pp. 138–151.

Bojanc, R., Jerman-Blažič, B. (2008) An economic modelling approach to information security risk management. *International Journal of Information Management, 28*(5), pp. 413–422.

Chang, S. E. in Ho, C. B. (2006) Organizational factors to the effectiveness of implementing information security management. *Industrial Management & Data Systems, 106*(3), pp. 345–361.

Data breach investigation report. (2014) New York: Verizon. Accessed 20. 7. 2014 on http:// www.verizonenterprise.com/DBIR/2013/.

Defending yesterday: Key findings from the global state of information security survey 2014. (2014) London: Price Waterhouse Coopers. Accessed 22. 7. 2014 on http://www.pwc.com/gx/en/consulting-services/information-security-survey/download.jhtml.

Feng, N. et al. (2014). A security risk analysis model for information systems: Casual relationships of risk factors and vulnerability propagation analysis. *Information Sciences, 256*, pp. 57–73.

Global corporate IT security risks. (2013). Moscow: Kaspersky Lab. Accessed 15. 6. 2014 on http://media.kaspersky.com/en/businesssecurity/Kaspersky_Global_IT_Security_Risks_Survey_report_Eng_final.pdf.

Goel, S., Shawky, H. A. (2009). Estimating the market impact of security breach announcements on firm values. *Information & Management, 46*(7), pp. 404–410.

Internet security threat report. (2014). Volume 19. Mountain View: Symantec Corporation. Accessed 20. 8. 2014 on http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf.

Kshetri, N. (2006). The simple economics of cybercrime. *IEEE Security and Privacy, 4*(1), pp. 33–39.

Mishra, S., Chasalow, L. (2011). Information security effectiveness: A research framework. *Issues in Information Systems, 12*(1), pp. 246–255.

Pironti, J. P. (2007). Developing metrics for effective information security governance. *ISACA Journal, 7*(2), pp. 1–5. Accessed 20. 7. 2014 on http://www.isaca.org/Journal/Past-Issues/2007/Volume-2/Pages/Developing-Metrics-for-Effective-Information-Security-Governance1.aspx.

Ponemon Institute. (2010). *Security effectiveness framework study*. Traverse city, MI: Ponemon Institute. Accessed 1. 6. 2014 on http://h71028.www7.hp.com/enterprise/downloads/software/Security%20Effectiveness%20Framework%20Study.pdf.

Rhee, H. S. et al. (2012). Unrealistic optimism on information security management. *Computers & Security, 31*(2), pp. 221–232.

Saleh, S. (2006). A new approach for assessing the maturity of Information Security. *ISACA Journal, 6*(3), pp. 1–7. Accessed 20. 7. 2014 on http://www.isaca.org/Journal/Past-Issues/2006/Volume-3/Documents/jpdf0603-A-New-Approach.pdf.

SANS Institute. (2013). *Critical security controls, version 5*. Accessed 19. 8. 2014 on http://www.sans.org/critical-security-controls.

Son, J. Y. (2011). Out of fear or desire? Toward a better understanding of employees' motivation to follow IS security policies. *Information & Management, 48*(7), pp. 286–302.

Stewart, A. (2012). Can spending on information security be justified? *Information Management & Computer Security, 20*(4), pp. 312–326.

The impact of cybercrime on business. (2012). Traverse City: Ponemon Institute. Accessed 20. 7. 2014 on http://www.checkpoint.com/products/downloads/whitepapers/ponemon-cybercrime-2012.pdf.

Detecting and Correlating Supranational Threats for Critical Infrastructures

Konstantin Böttinger, Gerhard Hansch and Bartol Filipovic
Fraunhofer AISEC, Garching near Munich, Germany
Konstantin.Boettinger@aisec.fraunhofer.de
Gerhard.Hansch@aisec.fraunhofer.de
Bartol.Filipovic@aisec.fraunhofer.de

Abstract: As critical infrastructures have become strategic targets for advanced cyber-attacks, we face the severe challenge to provide new defense technologies for their protection. We propose a distributed supranational architecture for detection, classification, and mitigation of highly sophisticated cyber incidents targeted simultaneously at multiple critical infrastructures. We build upon a three layered architecture comprised of Security Operations Centres at organizational (O-SOC), national (N-SOC), and European (E-SOC) level using IDS and SIEM solutions. In our approach we combine machine learning and automatic ontological reasoning: First, we apply methods from the field of machine learning to analyse threat indicators of different granularity. This provides classification of very specific observables collected at compromised sites. Second, we perform ontological analysis to identify large scale correlations within an incident knowledge graph. This yields insight into ongoing attack campaigns, especially regarding extent and expected impact. Our approach further allows to identify targets that are likely also to be affected or already compromised. Our proposed architecture counters advanced threats targeted against the critical infrastructures of Europe. We currently develop a prototype of our approach within the framework of European Union FP7 project ECOSSIAN (607577).

Keywords: threat detection, machine learning, ontological reasoning, critical infrastructures

1. Introduction

The increasing levels of interconnection and interdependency within critical infrastructures (CI) have significantly enlarged their attack surface. Disruptions of these vital systems such as energy, communication, or transportation may substantially impact on modern society. While there has been research effort for protection of isolated domains within single nation states, it is still an open question how to effectively implement a pan-European incident management system handling threat detection, large scale attack correlation, and early warning. This becomes even more challenging in the light of multistage attacks exploiting interdependencies of multiple CIs. Advanced threats such as Stuxnet, Duqu, and Flame are on the rise while the extent and full impact of future similar attacks are not predictable at present time. Only if we perform threat detection on a pan-European level we will gain full situational awareness regarding cyber incidents.

Threat information highly varies in granularity and completeness. Noisy and low-level threat information gathered within single sites provides detailed local information but misses the relationship to the overall threat situation. In contrast large-scale correlation technologies are not suited to process the vast amount of detailed and noisy threat information data collected locally. To fit local information into the overall context, we combine machine learning classification for low-level data and ontologic reasoning. This facilitates situational awareness, early-warning to possibly affected critical infrastructures and optimal mitigation strategies.

In summary, we make the following contributions:

- We propose a threat detection and correlation approach combining machine learning for fine-grained low-level information classification and ontologic reasoning for large-scale threat correlation.
- We define a distributed architecture for managing the complex analysis processes conducted simultaneously at multiple sites.
- We define information flows and structures that guarantee efficient processing and fast response in case of an attack.

The remainder of this paper is organized as follows. After the discussion of related work in Section 2 we present a motivating example in Section 3 to illustrate the problem setting. Features for threat detection and correlation are highly diverse (Section 4) and need to be harmonized (Section 5). In Section 6 we introduce our threat correlation method. Subsequent to the architecture and data flows in Section 7, we conclude with a discussion and outlook in Section 8.

2. Related work

Our method relates to technical publications and integrates into the policy framework of official authorities on EU level:

Anomaly detection and classification within industrial control systems have been studied by e.g. (Mantere, et al., 2013), (Pleijsier, 2013), and (Slot & Kargl, 2015). These approaches detect local incidents of isolated sites, but do not relate the results to remote (previous or current) impairment on a large scale. There is a vast amount of behaviour-based malware clustering efforts and we refer to (Sharif, et al., 2009) and (Bayer, et al., 2009). Regarding ontologic reasoning on high-level threat information, we use approaches presented by (Martimiano & dos Santos Moreira, 2006). To build knowledge graphs we combine different concepts for security modelling and analysis (Kim, et al., 2005), (Fenz & Ekelhart, 2009), system resilience (Vlacheas, et al., 2011), and incident management (Mundie, et al., 2014).

We build upon a three-layered architecture that enables a seamless integration into already existing CERT facilities and common procedures. An architecture comprised of Security Operations Centres at organizational level (O-SOC), national level (N-SOC), and European level (E-SOC) was discussed by (Kaufmann, Hutter, Skopik, & Mantere, 2015). There are significant efforts by ENISA and the European Commission related to pan-European CI protection and incident handling, e.g. (Bronk, et al., 2006), (ENISA, 2013), (ENISA, 2015), (Deloitte Bedrijfsrevisoren; ENISA, 2015), and (European Commission, 2013). Further, the NIST has identified major SCADA-related security aspects in (National Institute of Standards, et al., 2014). Our approach is developed within the framework of European Union FP7 project ECOSSIAN (607577) (ECOSSIAN, 2014) where several components directly connected to our architecture are being researched. Further, the VIKING FP7 project provided valuable insight regarding SCADA security.

3. Example scenario

This section provides a short scenario to illustrate our approach. We build upon information from CI operators directly represented within ECOSSIAN. For further use cases on gas distribution providers, infrastructure operators, financial service operators, and supply chains we refer to the ECOSSIAN deliverable D1.5 - Use Case Scenario Report.

We assume the objective of an attacker to be a disturbance in the field levels of an operator. Such attacks on the physical part of a cyber-physical system are complex in nature and we refer to (Gollmann, et al., 2015) for an example attack targeted against a vinyl acetate monomer plant. In order to achieve his goal, the attacker starts with infiltrating systems at the ERP level of the operator and constructs a pathway down to the field level. A typical entry point is the infection of a single PC in the office network using spear phishing. From there the attacker exploits a vulnerability in an update distribution server which holds the links for software updates requested by the MES and SCADA systems. The attacker further injects a specially crafted malware into the central server that receives command and controls via a DNS-tunnel. By redirecting update requests from the MES and SCADA systems, malicious firmware is deployed to the PLCs at the control level. Subsequently, the attacker is able to manipulate the behaviour of production facilities in the field according to his objectives.

Both fine- and coarse-grained features (see Section 4) are monitored at the different layers of the operator. First, the monitoring module at the sensor level detects an anomaly in measured values transferred to the actuators. Certain critical valves controlling the supply of gas within the targeted zone on field level periodically show dysfunction. We refer to the monitoring values as sample S1.

During incident analysis, a malicious file sample S2 is found on a SCADA server at the process level. Further investigation shows that the file must have come from the compromised update distribution server at the ERP level. The SIEM system at the ERP level detects a file sample S3 at the application server and further captures network traffic for one day. We refer to this network recording as sample S4. The security analyst finally identifies the entry point, i.e. the compromised PC in the office network. She generates a report S5 indicating details of the spear phishing attack vector, such as time of corruption, level of sophistication, and PC system information (e.g. operating system and infected applications).

The samples S1 to S5 differ with regard to their information content and the granularity of features: While S1 holds fine-grained control and sensor values, S5 contains textual information such as the initial attack vector. To

correlate the detected incident to other attacks we first have to fit the information and all IoC to processable classes within our hierarchy schema. First we map all low-level features to classes such as "*critical dysfunction by periodic overdrive"* for S1 or *"communication via DNS tunnel"* for S4. The first building block of our approach uses classification methods from the field of machine learning to assign the fine-grained and often noisy features to predefined labels. The identified classes, together with the coarse grained features (from e.g. S5) are then given to a correlation module which performs large-scale – EU-wide – ontologic reasoning on a shared knowledge graph.

Hence the reasoning module is able to compile

- all incidents found in the same domain (e.g. gas supply nets) in the last month within Europe,
- all incidents revealing the same attack sequence (S1 to S5), or
- a list of operators (e.g. gas power plants) that depend on the supply net of the targeted operator in order to cast an early warning.

4. Feature diversity

In this section we exemplary highlight the diversity of fine-grained and coarse-grained features for threat analysis. Potential feature sources are described in (Cichonski, Millar, Grance, & Scarfone, 2012).

4.1 Fine-Grained features

There is a multitude of low-level features that would be suited for classification using machine learning. For illustration, we present some possible features here, but note that the features for a real-world implementation need to be selected according to evaluation of the running detection system. Feature selection will then be provided by experiment with the deployed and running threat detection system.

Non-Executable File Features: Features for classification of non-executable files may include meta-data as well as detailed characteristics of the file. The choice of features strongly depends on the specific deployment of the classification module. For example, if the malicious file is an instruction list for a Programmable Logic Controller (PLC) in the field level, features could be identifier, input, mode, time basis, programmed value, actual value, and modifiable flag of specific sections.

Executable File Features: File features for executable files may include characteristic features such as Application Binary Interface (ABI), system calls, type of subroutines (File I/O, Threat, Network, GUI, Registry), number of branches in specific sections, exported functions, or mutexes.

Network Traffic Samples: Features of network captures may include information about packet sizes, throughput, session frequency, flow direction (depending on the initiating system of the communication), protocol and protocol settings, and entropy of messages. Mantere et al. (Mantere, Sailio, & Noponen, 2013) discuss and evaluate further possible network features in industrial control systems and we refer to their paper for a comprehensive overview.

Sensor/Actor Features: Typical features for sensor and actor communication are control message frequency and set values (e.g. temperature values in a predefined time, liquid level, or voltage).

4.2 Coarse-Grained features

Similar to low-level feature selection, there is a variety of possible choices for high-level features we can choose from for the reasoning. The following features are exemplary chosen and shall give an impression about the various possibilities. We basically refer to meta-data of the infection as well as environment data. The selection is motivated by (Slot & Kargl, 2015).

Attack Vector and Line of Action: Based on the incident analysis, it is important to reconstruct the initial attack vector as part of the overall line of action the attacker used. While most incident reports provide only the initial attack vector of an attack, the exact line of action supports the attacker recognition as they often follow an individual procedure (same moving pattern, tools and methods) to perform their tasks across different victims.

Infection and Stealth Mechanisms: We can further take into account infection and stealth mechanisms such as MBR infection, hiding in partitions, or interrupt and message hooks. Further features for large-scale correlation are encrypted network communication, number of distinct IP addresses, steganographic capabilities (e.g. DNS tunnelling), and fast-flux exfiltration.

Injection Targets: Injection targets may include specific servers and services, infected databases, targeted sensor and actor components, and network infrastructure (e.g. connection to ERP systems).

Target Environment: The target environment features of an attack may include information about the domain and zone where it is installed and operates, the connectivity, and the surrounding IT infrastructure.

Time and Periodicity: Attacks often follow a characteristic timing behaviour. Malicious programs for example communicate to their command and control servers in predefined time frames. The periodicity and timing can reveal valuable unique patterns. Another example is given by the timing of multi-stage attacks. If there are several local IoC in a row, the timing of such a sequence may reveal a certain already known threat pattern.

Impact: The impact features include information about the local impact on the affected machine(s) like a loss of functionality, manipulation of processed or displayed values, remote control, malware spreading, or a permanent destruction of the component. The observation of several known IoC at different devices within one group may also reveal an already known threat pattern.

Interdependencies to Critical Infrastructures: This feature includes information about relations of parts or services other CIs depend on. This is necessary to correlate attacks and identify the direct and indirect affected victims. It is especially important as it is possible that the attack was just a single step of a large-scale campaign with the aim to disturb the operation of another CI by breaking its supply chain.

5. Feature harmonization

How can we harmonize the diversity of monitored data in a way that allows correlation and ontologic reasoning without information loss? To get overall situational awareness we have to fuse data from low-level sources with abstract and general environment data. For example, if a malicious executable is found by a locally deployed SIEM and we want to fit this local information into the pan-European context, we have to combine fine-grained information about the malicious executable with high-level information data: Where has this file has been sighted in the past? Under which facility environment? Which mitigation strategies were applied in reaction? In a twofold approach for threat information analysis, we first process and classify low-level data locally with methods from the field of machine learning. Second, we apply ontologic reasoning to understand how to integrate the output of this classification into the overall threat landscape.

Classifying a given sample means assigning predefined labels. If there are multiple labels to be assigned to the sample, we refer to the approach as multi-label classification, and if there are more than two possible classes that can be assigned to the sample we apply multiclass classification methods. In a supervised learning setting, the classifier is given a set of training samples together with corresponding labels. In the learning step, the classifier processes this training set to adapt its parameters. When given an unknown sample not in the training sample set, the trained classifier assigns a label to the sample based on the preceding training step. This way we can classify unknown and noisy samples.

There is a diversity of classification methods that suit our purposes of handling noisy fine-grained features, e.g. linear classifiers, support vector machines, kernel methods, neural networks, decision trees, or random forests. Regardless of which method is applied, the classifier takes as input a set of fine-grained features and a sample threat to be classified and outputs a set of classes. Such a mapping from fine-grained features to class labels provides us with a common level of abstraction for the indicators of compromise which is necessary for subsequent ontologic reasoning. Analogous to the classification we performed in our initial example (see Section 3) we can think of other class labels such as malware family identifier, threat target, communication mechanisms, and threat actors. In general, classification of low level features of suspicious files or network traffic is capable of detecting labels of a predefined set. The detected classes are then redirected to the ontologic reasoning module.

6. Threat correlation

The threat labels gained from fine-grained feature classification are now on a comparable level of abstraction with other indicators of compromise gathered from SIEM and IDS solutions. This enables us to correlate them to previous threats using ontological reasoning.

An ontology includes definitions of the classes, properties, and their relationships. It serves as a base to store information in a machine-readable knowledge graph. Statements about resources are typically formatted using the Resource Description Framework – RDF (Klyne, Carroll, & McBride, 2014), while the Resource Description Framework Schema (RDFS) (Brickley, Guha, & McBride, 2014) is used as a (hierarchical) class schema. Domains that make heavy use of ontologies are e.g. medicine (W.H.O., 2015), linked web data (LOV, 2016) and knowledge organization systems. Further, the use of a description logic puts a system in a position to make formal qualified assumptions. The Web Ontology Language OWL (Motik, Patel-Schneider, & Horrocks, 2006) in its decidable description logic variants (e.g. OWL-DL) is used to support the creation of formal correct descriptions, processing, and understanding. Although a separate program is able to draw conclusions despite heterogeneously grained information, it is more efficient and less prone to errors to use a semantic reasoner like HermiT (Glimm, Horrocks, Motik, Stoilos, & Wang, 2014) or Pellet (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007). Thus, conclusions can be drawn by the use of ontologies with a determined taxonomy and semantic relations enabling entailment relations and request output by a semantic query language like SPARQL (Prud'hommeaux & Seaborne, 2013).

To correlate threats, we need to aggregate and process coarse-grained features as presented in Section 4.2. *Knowledge graphs* archive and relate the detected class labels (of Section 5), coarse-grained feature vectors, and IoC, which allows us to formulate a diversity of queries to search for correlations and patterns.

The *Global Knowledge Graph (GKG)* is hosted and maintained by the E-SOC and serves as a growing repository of technical information about incidents, samples, coarse- and fine-grained features, and the relations between them. All information is stored in a unified manner using a hierarchical incident ontology, extended with IoC and system feature classification schemes. The GKG serves search requests from each SOC where required. The identity of the information source (the affected operator) and all related references are stored pseudonymized and only get disclosed when a substantial interest is justified.

By accessing the GKG, previously identified attack patterns are recognised, correlated and conclusions about the expected attacker behaviour and their final target can be drawn. For example, a CI operator under attack can send requests to the GKG regarding the expected next steps of the attacker, the impact to be expected, or the potential final aim of the attack.

An individual *Local Knowledge Graph (LKG)* is hosted and maintained by each SOC. It serves as a memory for restricted or sensitive individual information like detailed system descriptions and configurations, business dependencies, and organizational structures. N-SOCs and E-SOC particularly are aware of interdependencies of supervised CIs and know their respective technical services. While direct relations between two operators are rather obvious, it becomes complex to identify linkages across a whole supply chain. This is because dependencies between operators can appear not only in directed forms like supplier-consumer or service-provider, but also include cases where two or more operators depend on each other. Therefore, it is necessary to track and solve interdependency information across all SOC layers.

By using machine learning technologies to fit information into an incident management ontology like CIMBOK (Mundie, et al., 2014), ENISA Ontology (Vlacheas, et al., 2011) and ONTOSEC (Fenz & Ekelhart, 2009), we receive consistent relations and classes that enable an automated analysis. Further, implicit information in the knowledge graph is made explicit by inference reasoning and can then be retrieved using a semantic query language. Automated reasoning results trigger CSIRTs on incidents that would otherwise stay undetected, like e.g. accumulation of incidents sharing a minor common feature.

Our solution includes different threat detection and correlation steps to be performed at each SOC level. The necessary information is exchanged using STIX (MITRE Corporation, 2015) messages in combination with RDF/XML triples in the common ontology.

In order to derive context information, the ontologic reasoning module performs an inference from the detected features and the information from the LKG and GKG. During all phases of the analysis each new information is submitted to the N-SOC, including all IoC, classes, performed measures, and information about the affected systems and services. This processing enables a more efficient local incident handling compared to a sole CSIRT that tries to handle the actual situation.

The correlation module supports the N-SOC with correlation of incidents on a national level, identification of expected impacts, and issuing of specific warnings to CI operators. To find common patterns between distributed incidents, the features labelled by the classifier and reported by the O-SOCs are matched with those of prior and current situations using the semantic reasoner. By correlating the received incident notifications from multiple O-SOCs, the N-SOC can gain insight regarding the severity and extent of the campaign, other CIs that might also be vulnerable to the attack (as they operate similar installations), and reveal hints about possible attackers. Operators of CIs that are assumed to be vulnerable to the attack are immediately alerted by the N-SOC to watch out for the found IoC and take according actions like described by (Foley & Fitzgerald, 2011). Beside the incident management support, the N-SOC also serves as a filter for messages between the E-SOC and the O-SOCs. It forwards information about features, classes, and relations detected by the O-SOC to be added to the GKG by the E-SOC. During this process, it generalizes sensitive features such that they reveal no sensitive business information about the O-SOC. In the other direction, the N-SOC forwards warnings from the E-SOC to the O-SOCs where required.

Further, the correlation module supports the E-SOC on its task to monitor and coordinate the activities of the N-SOCs. Therefore, it supports the detection of (large-scale) attack campaigns and the issuing of specific warnings to CI operators. Additionally, the correlation module resolves dependencies between different CI domains. As the N-SOCs submit only non-sensitive and generalized (the classes of sensitive context) information due to privacy reasons, the E-SOC operates on coarser grained information. The E-SOC searches large-scale attack indicators by correlating input from the N-SOCs, the LKG, as well as the common GKG. Thereby, accumulations of striking patterns (domains, areas, timings, etc.) as well as targeted supply chains can be detected. The STIX alert messages sent from the E-SOC to the N-SOC include a list of classes and IoC that are assumed to be at risk. This leads to a situation specific early warning for operators of similar or dependent systems across borders without revealing sensitive information about treat or operators.

7. Architecture and data flows

The overall architecture of our correlation module (CM) is depicted in Figure 1: Architecture of the threat detection module Monitoring modules (M) of each CI layer forward the recorded samples to the feature aggregator (FA). Coarse-grained feature values (as described in Section 4.2) are directly extracted and forwarded to the ontologic reasoning (OR) module. Fine-grained features are first classified in the fine-grained feature classification (FFC) module (as described in Section 4.1). The resulting class labels are then forwarded to the OR module. The coarse-grained features values as well as the class labels for the reported incident are then correlated to the LKG and the global GKG (see Section 6).

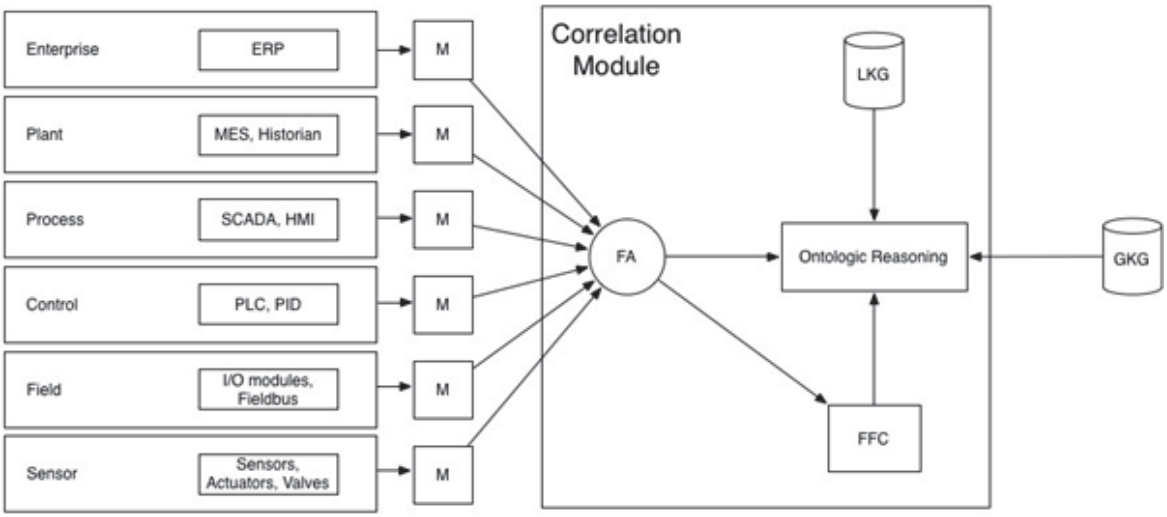


Figure 1: Architecture of the threat detection module

As depicted in Figure 2: Communication and access structure, each SOC deploys its own CM. The ontologic reasoning modules of all correlation modules share one GKG. For a practical implementation this GKG will be mirrored to provide backup and recovery in order to minimize the risk of a potential outage of this central component.

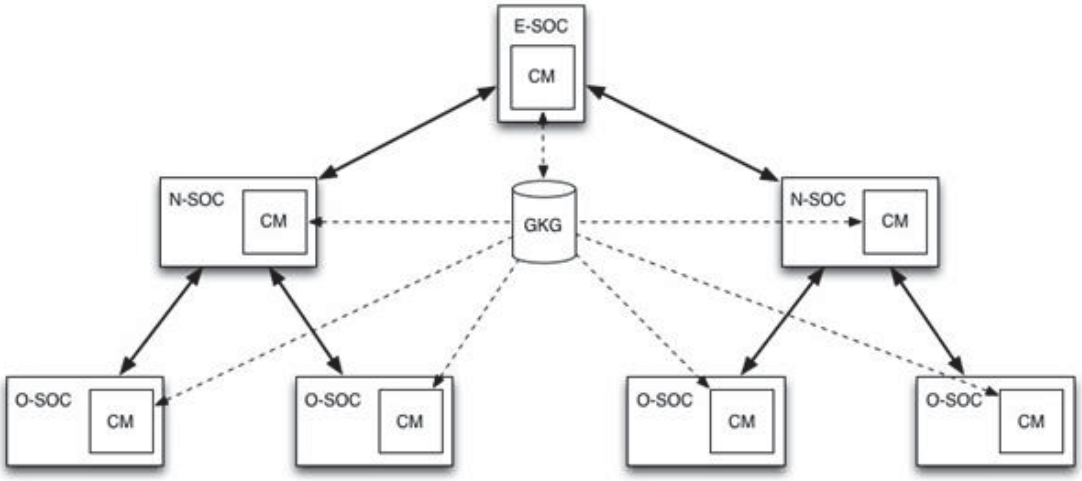


Figure 2: Communication and access structure

The O-SOC CMs forward suspicious as well as confirmed IoC to the respective N-SOCs, which take the received data to perform correlation of national threats. The N-SOCs then forward correlation results to both to the E-SOC and related O-SOCs. In case the FFC of an O-SOC cannot label a given sample to a predefined class, it forwards the monitored sample to the N-SOC. Similarly, if the N-SOCs FFC cannot find a fitting label, it forwards the sample to the E-SOC. This way the N-SOCs and the E-SOC are capable of finding and defining new class labels for anomaly patterns that are unknown to the O-SOCs. The O-SOCs FFC parameters for classification and class label lists are regularly updated with the findings of the N-SOC, and analogue the N-SOCs FFCs receive updates from the E-SOC.

8. Discussion and conclusion

We present a distributed architecture for detection, classification, and correlation of multi-stage attacks targeted against the EU's critical infrastructures. Such a correlation system has to solve the problem of diverse information. The granularity of possible indicators of compromise ranges from low-level signals from the field level of an operator to an incident report generated by a security analyst. We propose a method to define a common level of abstraction by transforming fine-grained features into coarse-grained classes. This classification pre-processing allows correlating threat information to previous and supranational attack patterns stored in ontological knowledge graphs. Correlation is done at operational, national, and European level and we define local and global knowledge graphs in order to provide separation of sensitive information and sharing of correlation insights with potentially affected operators. While classification is provided by methods from the field of machine learning, we use ontologic reasoning to correlate and detect new attack patterns. The knowledge graphs store indicators of compromise in RDF format which enables us to formulate a diversity of queries in order to search for correlations and patterns. As new correlation insights are directly propagated to the N-SOCs and E-SOC, our system provides a basis to cast early warnings to dependent CIs. We currently develop a prototype of our approach within ECOSSIAN.

Acknowledgements

This research is partially supported by the framework of European Union FP7 project ECOSSIAN (607577).

References

Brickley, D., Guha, R. & McBride, B., 2014. *RDF Schema 1.1.* [Online] Available at: https://www.w3.org/TR/rdf-schema/

Bronk, H., Thorbruegge, M. & Hakkaja, M., 2006. *A STEP-BY-STEP APPROACH ON HOW TO SET UP A CSIRT.* [Online] Available at: https://www.enisa.europa.eu/activities/cert/support/guide

Cichonski, P., Millar, T., Grance, T. & Scarfone, K., 2012. *Computer Security Incident Handling Guide SP 800-61 Revision 2.* National Institute of Standards and Technology.

Deloitte Bedrijfsrevisoren; ENISA, 2015. *Cyber Security Information Sharing: An Overview of Regulatory and Non-regulatory Approaches.* [Online]

ECOSSIAN, 2014. *FP7 Security Project ECOSSIAN - European COntrol System Security Incident Analysis Network.* [Online] Available at: http://www.ecossian.eu/

ENISA, 2013. *Detect, SHARE, Protect.* [Online] Available at: https://www.enisa.europa.eu/activities/cert/support/information-sharing/detect-share-protect-solutions-for-improving-threat-data-exchange-among-certs

ENISA, 2015. *Incident Handling Automation.* [Online] Available at: https://www.enisa.europa.eu/activities/cert/support/incident-handling-automation

European Commission, 2013. *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning measures to ensure a high common level of network and information security across the Union*

Fenz, S. & Ekelhart, A., 2009. Formalizing information security knowledge. p. 183–194.

Foley, S. N. & Fitzgerald, W. M., 2011. *Management of security policy configuration using a Semantic Threat Graph approach.*

Glimm, B. et al., 2014. HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning*, 10, 53(3), p. 245–269.

Gollmann, D. et al., 2015. Cyber-Physical Systems Security. p. 1–12.

Kaufmann, H., Hutter, R., Skopik, F. & Mantere, M., 2015. A structural design for a pan-European early warning system for critical infrastructures. *e & i Elektrotechnik und Informationstechnik*, 132(2), p. 117–121.

Kim, A., Luo, J. & Kang, M., 2005. *Security ontology for annotating resources.* Springer.

Klyne, G., Carroll, J. J. & McBride, B., 2014. *RDF 1.1 Concepts and Abstract Syntax.* [Online] Available at: https://www.w3.org/TR/rdf11-concepts/LOV, 2016. *Linked Open Vocabularies.* [Online] Available at: lov.okfn.org/dataset/lov/Mantere, M., Sailio, M. & Noponen, S., 2013. Network Traffic Features for Anomaly Detection in Specific Industrial Control System Network. *Future Internet*, 5(4), p. 460–473.

Martimiano, L. A. F. & dos Santos Moreira, E., 2006. The Evaluation Process of a Computer Security Incident Ontology.

MITRE Corporation, 2015. *Structured Threat Information eXpression (STIX™).* [Online] Available at: http://stixproject.github.io/

Motik, B., Patel-Schneider, P. F. & Horrocks, I., 2006. *OWL 1.1 Web Ontology Language Structural Specification and Functional-Style Syntax.* [Online] Available at: https://www.w3.org/Submission/owl11-owl_specification/

Mundie, D. et al., 2014. An Incident Management Ontology.

National Institute of Standards, Technology & United States of America, 2014. Framework for Improving Critical Infrastructure Cybersecurity.

Pleijsier, E., 2013. Towards anomaly detection in SCADA networks using connection patterns. p. 1–6.

Prud'hommeaux, E. & Seaborne, A., 2013. *SPARQL Query Language for RDF.* [Online] Available at: https://www.w3.org/TR/rdf-sparql-query/

Sirin, E. et al., 2007. Pellet: A practical OWL-DL reasoner. *Software Engineering and the Semantic Web*, 5(2), p. 51–53.

Slot, T. & Kargl, F., 2015. Detection of APT Malware through External and Internal Network Traffic Correlation.

Vlacheas, P. T. et al., 2011. *Ontology and taxonomies of resilience.*

W.H.O., 2015. *International Classification of Diseases (ICD).* [Online] Available at: http://www.who.int/classifications/icd/en/

A Cross-Disciplinary Approach to Modelling and Expressing Adversity

Ian Bryant, Carsten Maple and Tim Watson
Cyber Security Centre, WMG, University of Warwick, UK
i.bryant@warwick.ac.uk
c.maple@warwick.ac.uk
t.watson@warwick.ac.uk

Abstract: There is a significant, unmet requirement for risk and hazard (jointly referred to as adversities) enumeration methods that are applicable across currently stovepiped domains, replicable and scalable in approach. This paper considers the issues relating to replicable and scalable methods for the enumeration of adversities, and proposes a method that provides an understanding of all categories of risk, not only those arising from the actions of cyber antagonists, but also from factors such as natural disasters. The method provides an approach to modelling adversity in a way that combines differing types of adversity, an absolute scale for the enumeration of adversity impact, a summary statistic model and plot that reflect the probabilistic nature of adversities, and the Annualised Expectation of Risk (AER) as a way to facilitate common interpretation of adversities across and within differing communities of interest.

Keywords: risk, hazard, adversity, risk enumeration, risk measurement, risk aggregation, risk management

1. Introduction

Given the lack of a general, agreed definition and approach to risk identification, risk measurement and risk management (Aven 2012A) it can be argued that the broad understanding of risk (as opposed to the narrow understanding of discipline-specific aspects of certain types of risk) is clouded by the preponderance of 'stovepipes' – self-organised confederations of experts and practitioners – and by the questionable validity of many quasi-numerical approaches to risk quantification.

This has always been a problem, but the fast evolving, complex, dynamic nature of the cyber environment provides a potentially disruptive challenge to existing views and methodologies. This is further compounded by the intrinsic degree of interconnectedness and dependence both within the cyber environment itself and by the emerging degree of interconnectedness and dependence within the physical world; the implied separation between the cyber and physical domains is, in itself, yet another stovepipe. Consequently, it appears that there is a significant, unmet requirement for risk enumeration methods that are applicable across currently stovepiped domains/disciplines that are also replicable and scalable in approach. This paper outlines one promising method that provides a better understanding of all categories of risks and hazards across multiple communities of interest. While other approaches have been suggested previously (e.g. Gardoni and Murphy 2014), these have still been quite narrowly focused to specific disciplines or domains. A contribution of the method outlined in this paper is that it is generally applicable across all disciplines and domains

This paper discusses the management of risks in Section 2 and presents a taxonomical approach in Section 3. A measurement approach and an expression approach are presented in Sections 4 and 5 respectively, before the Risk Expectation is introduced in Section 6. The paper concludes with Section 7 presenting both conclusions and recommendations for further work.

2. Management of risks

2.1 Risks

Any and all entities, regardless of their size or sector, have a common need to seek to properly manage their risks i.e. to identify probable challenges, use appropriate countermeasures to reduce exposure to an acceptable level, and have a structure for dealing with new risks that emerge (ISO/IEC 31000(2009))

The approach to Risk Management as advocated in the ISO/IEC principles and other good practice is that before such risks can be managed, they first need to be described and/or enumerated. Although the detail as to how such description and/or enumeration is performed varies between domains and communities of practice (Aven 2012B), in particular in terms of the terminology used, the five main factors can be generalised as:

- The set of Assets to be protected
- The set of Adversities that are faced (with "Threats" or "Hazards" being the most common subsets)
- The set of Compromise Paths that are exposed
- The Risk Appetite of the entity
- The set of Controls that are available for application

2.2 Stovepipes

Within the cyber domain, there are predictably a plethora of stovepipes. Two different levels of stovepipe, varying in granularity, are particularly noteworthy in the context of Risk. Firstly, there is a set of high stovepipes predicated upon the organisational role performed by a practitioner, with three primary such clusters in the cyber domain being:

- The Business Stovepipe, with a focus on the effects of all operational risks as they impact on organisational outcomes
- The ICT Delivery Stovepipe, with a focus on standard Project and/or Business As Usual (BAU) "SQIRT" risks (Scope, Quality, Resource, Implementability, Timescale)
- The Protection Stovepipe, with a focus on the risks from operational adversities

This has the consequence that risk will be used in at least three differing and sometimes partially incompatible manners. Stovepipes further exist at lower levels of granularity. For example, consider the Protection Stovepipe, where there is additional stovepiping a this lower level of granularity:

- A threat-centric stovepipe – typically called security – that is concerned about actions of cyber antagonists
- A hazard-centric stovepipe – typically called safety – that is concerned about factors such as natural disasters

Yet the consumer of cyber services, be that for the data (for Operational Technologies – OT) or the information (for Information Technologies – IT, and Consumer Technologies – CT) that is stored, processed or forwarded, or the technologies which perform the storage, processing or forwarding, is actually not interested in such distinctions – it matters little to them whether a disruption to IOCT (Information, Operational, Consumer Technologies) results from a hazard or a threat. As a consequence, it seems sensible to treat both hazards and threats together as one joint concept: adversities.

2.3 Adversities

The concept of adversity has been proposed previously (Bryant 2012) to address this perceived, yet artificial stovepipe distinction between the views of the security and safety communities, in which the security community seeks to address threats (directed, deliberate, hostile acts) and the safety community seek to address hazards (undirected events). Although there is a logical distinction between the two (a threat is normally viewed as being the result of a human actor who has intent/motivation and capability), an abstraction can be usefully taken to combine them as a superset, adversities, with associated probabilities of occurrence.

2.4 Confounding factors

A challenge to the modelling of risks is the role of uncertainty in their enumeration, with a four-way delineation having been proposed (King 1989), as shown in Table 1. A simpler approach was popularised as the Known-unknown-Unknowable (KuU) model (Gomory 1995) which subdivides risks into three categories, which can be mapped back to the original work by King as follows:

- Known [K: Type I/II] - an approximation to both frequency and magnitude can be made as a probability distribution
- Unknown [u: Type III] - an approximation to both frequency and magnitude can be made as a probability distribution
- Unknowable [U: Type IV] - neither frequency nor magnitude can be sufficiently characterised

Table 1: Risk types

Risk Type	Event Nature	Outcome Probabilities	Characteristic
I	Known	Fixed	Deterministic
II	Known	Known	Stochastic
III	Known	Uncertain	Uncertainty
IV	Unknown	Unknown	Emergence

Using the KuU model reveals an underlying methodological issue in taking a holistic approach to adversity, as the differing stovepipes tend to differing approaches. The security world assumes a deterministic threat model, which typically ignores hazards, and is largely predicated upon characterisation of known types (if not necessarily details) of threat actor (Cox 2008), which therefore has difficulties handling the full KuU model. On the other hand, the safety community typically uses stochastic models to address hazards (Janich and Netjasova 2008), and usually ignores threats.

3. Taxonomical approach

A reasonable aspiration would be for those engaged in the delivery of cyber services to review and – as far as possible – manage all sources of adversity. Current modelling techniques have, however, largely aligned with the safety and security stovepipes, with scant attention being paid to how these differing viewpoints can be combined despite the obvious attractions of alignment having been identified (e.g. Brostoff and Sasse 2001, and Firesmith 2003).

3.1 Definitions

The Adversity Set (AS) approach (Bryant *et al*. 2014) is centred upon a taxonomical approach which allows the mapping of the variety of potential Adversity Factors (AF) into a common set of Adversity Classes (AC) to accommodate differing communities of interests' views in as domain- and community-neutral manner as possible.

We define **Equation 1**:

$$AC_y = \sum_{y}^{\cdot}(AF_{y.x})$$

where y = Adversity Factor name
x = Adversity Class name.

We define **Equation 2**:

$$AS = \sum_{y}^{\cdot}(AC_y)$$

where y = Adversity Class name

3.2 Modelling

An Adversity Modelling (AM) approach is proposed to produce a domain-neutral representation of adversities, which allows all AF to be captured, collated and simplified, to produce a set of combined Adversity Class that form the overall Adversity Set. An example of an Adversity Class would be kinetic adversity, which allows a number of different adversity factors that cause damage or disruption on a cyber or cyber–physical system to be combined into a single review point, including hazards (ranging from impact of varying categories of meteoroid, though wind-blown debris, to vehicle accidents) and threats (ranging from impact of a missile to an individual with a sledgehammer). The use of Adversity Model means that early stages of analysis will cause a proliferation of Adversity Factors to be considered, which while appearing contrary to the aims of improving usability, is crucial to simplify later stages.

Figure 1 is abstracted from a test adversity model Bryant and Watson (2014) that has been consciously selected to represent AF arising from both hazards and threats that, although largely and seemingly “non-technical”, can

nonetheless have significant impact on cyber or cyber–physical systems, and can clearly be seen to group into AC.

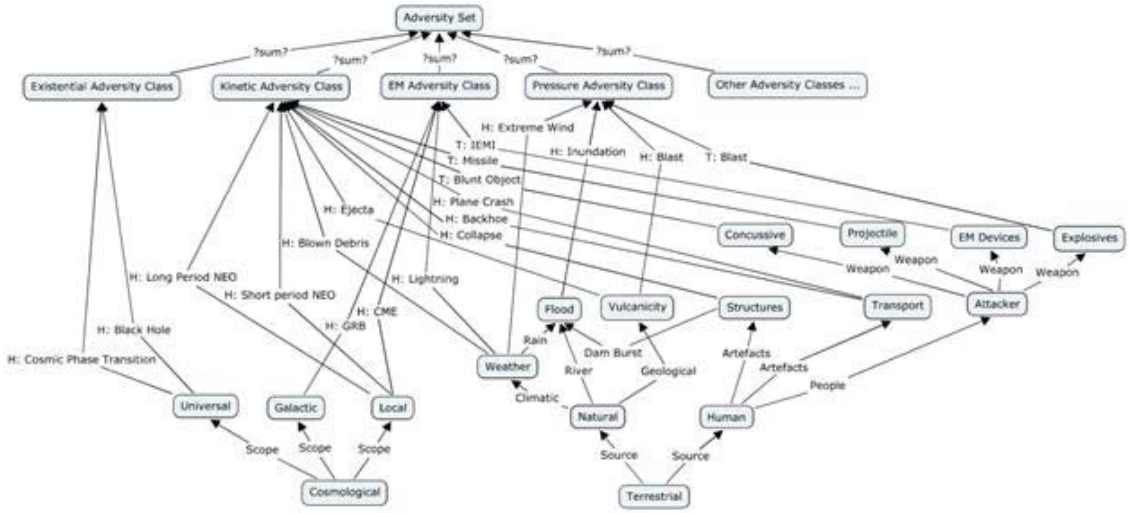


Figure 1: Example adversity model (partial abstract)

3.3 Taxonomy

A generic set of AC were derived initially from both academic literature (e.g. Baskerville and Im 2005) and other sources (e.g. Cabinet Office 2012) and then validated by exposure to a number of cyber-physical system scenarios. An updated version of the initial list (Bryant and Watson (2014) of generic Adversity Classes is provided in Table 2, reflecting the some minor changes of terminology .

Table 2: Adversity classes

Adversity Classes (AC)		Impact Class (IC)
AC.EX	Existential	IC.IOCT.Collateral
AC.KI	Kinetic Impact	IC.IOCT.Collateral
AC.PR	Pressure	IC.IOCT.Collateral
AC.IN	Inundation	IC.IOCT.Collateral
AC.TH	Thermal	IC.IOCT.Collateral
AC.CH	Chemical	IC.IOCT.Collateral
AC.BI	Biological	IC.IOCT.Collateral
AC.RD	Radiological	IC.IOCT.Collateral
AC.VA	Vanishment	IC.IOCT.Indirect
AC.TR	Trespass	IC.IOCT.Collateral
AC.EM	Electromagnetic	IC.IOCT.Indirect
AC.LD	Logical Disruption	IC.IOCT.Direct
AC.DD	Data Disruption (includes Destruction)	IC.IOCT.Direct
AC.DL	Data Leakage	IC.IOCT.Direct
AC.IM	Impede	IC.IOCT.Indirect
AC.FA	Failure	IC.IOCT.Indirect

3.4 Impact

As part of the validation exercise, the nature of impact was considered, with three impact classes being found to be relevant to IOCT as illustrated in Table 2:

- Direct – where the adversity acts directly upon the logic-bearing function or data/information within the subject IOCT

- Indirect - where the adversity acts directly upon the IOCT, but not directly on the logic-bearing function or data/information within the subject IOCT
- Collateral – where the adversity has an impact that impinges otherwise upon the IOCT or its logic-bearing function or data/information

4. Measurement approach

A measurement approach should to allow all relevant information to be considered, and be able to produce consistent results. Yet many current approaches do not perform well in terms of consistency (Hubbard 2009), and often have differing scales (Hubbard 2010).

4.1 Types of scale

Before proceeding with consideration of the needs of enumeration specific to adversity, a general understanding of Representational Measurement Theory (RMT) is needed, which tells us that measurements are often of differing natures (Stevens 1946 and Niederée 1992). A single measurement may be:

- Nominal – Qualitative Data produced by assigning observations into unranked categories
- Ordinal – Qualitative Data produced by assigning observations into ranked categories
- Interval – Quantitative Data produced by assigning ranked categories
- Ratio – Quantitative Data with an arbitrary baseline
- Absolute – Quantitative Data with a finite baseline

It should also be noted that when measurements are combined, the accuracy of any combined measurement will be no more accurate than the least accurate source datum.

4.2 Quantification of scale

In the context of an enumeration specific to adversity, an absolute scale would be preferable so that replicable results are produced. Many existing scales are based upon a monetary impact assessment, yet many of the practical impacts of adversity do not have a direct monetary value. A set of Deleterious Outcomes (DO) ranging from Regulatory Noncompliance thru' Economic Damage to Loss of Life (Bryant and Watson 2014) can be used as a conceptual framework from which to align monetary and non-monetary impacts, with "Value of Statistical Life" (VSL: also known as Value of Preventing a Fatality - VPF) based on a recent meta-analysis (OECD 2012) to establish a reasonable measure of central tendency for the most extreme of these Deleterious Outcomes. A similar concept exists of Value of Preventing Injury (VPI) but no comparable meta-analysis as yet exists, although several national jurisdictions intend to publish their own guidance (e.g. Judicial College 2013).

These anchors were used to produce a Deleterious Result Scale (DRS), which pegged the Deleterious Result (DR) of a fatality at the next highest rounded appropriate value in international currency units, of 1 Statistical Life = 2,000,000 XDR[1]. Such a value neatly illustrates the potentially large values that a DRS could assume, so to make the numbers more intelligible, a logarithmic, absolute scale is proposed, with 1 Statistical Life = DR7.0[2]. The scale has been extrapolated *ad absurdum* to illustrate what is arguably a finite concept of maximum risk with a DR of 41.18, albeit this would be challenged by Multiverse Hypothesis (Everett 1957).

The logarithm-based approach of the DRS has the collateral benefit of facilitating the instinctive filtration of less relevant risks due to the intrinsic order of magnitude steps in the characteristic of the DR[3].

From this baseline, a complete DRS enumeration can be derived, with Table 3 showing some major values.

1 ISO4217 Special Drawing Rights
2 Adjusted as (Log_{10} + 1.0) to allow 0 = 0.0
3 And also facilitates discarding excess detail by ignoring the mantissa

Table 3: Proposed deleterious result scale (DRS) (abbreviated)

DR	Preventable Fatalities	Example(s)	Economic Impact	
0.00	-	(No impact)	XDR	0
1.00	-	• Regulatory Noncompliance • Legal Offence • Disrupted Relationships • Disrupted Operations • Reputational Damage • Personal Distress • Economic Damage • Physical Damage • Personal Injury	XDR	2
2.00	-		XDR	20
3.00	-		XDR	200
4.00	-		XDR	2,000
5.00	-		XDR	20,000
6.00	-		XDR	200,000
7.00	1	One Person Dies	XDR	2,000,000
16.86	7.24E+09	World Populace on 20140701 Dies	XDR	1.45E+16
18.00	(1.00E+11)	(Notional loss of Solar System)	XDR	2.00E+17
29.48	(3.00E+22)	(Notional loss of Stars in Milky Way)	XDR	6.00E+29
41.18	(1.50E+34)	(Notional loss of Stars in Known Universe)	XDR	3.00E+40

4.3 Confounding factors

DR is an expression of likely impact, and as such matters are fundamentally uncertainties, it is important to remember that DR, although based on an absolute scale, is a replicable value but not a precise value. There is an overwhelming tendency to express and treat risks and their associated adversities as a single "magic number", which typically tends assume a simplistic, centre-tendency view of likelihood and uncertainty, or, in the limited number of cases where a probability distribution is assumed, this will typically tend to be modelled as a Gaussian (Savage 2002 and 2009). However, this view from both the producer (by implication) and consumer (by inference) is both naïve and unhelpful, especially as many cyber risks will have dramatically different underlying distributions:

- The large volume of Low Impact, High Probability (LIHP) adversities faced by many IOCT systems (e.g. network probes and malware infections) that for any individual instance would inherently count as an infinitesimal risk, yet the very existence of firewalls and anti-virus software (AVS) is a testament to the consensus that the aggregated risk is worth treating
- The underlying likelihood – which will typically be either unknown, or potentially even Unknowable – of new High Impact, Low Probability (HILP) adversities, a modern cyber-domain instantiation of the very essence of the Black Swan problem (Taleb 2007)

In addition to the problems of varying underlying distributions, the combined and/or blended nature of many real-world high-impact events (Perrow 1984) means that many cyber risks will be "Messes" (Ackoff 1974) or "Wicked" (Rittel and Kunz 1970). Such forms of Messes and Wicked risks are not always amenable to a systematic treatment, which leads to the needs to consider approaches such as the Soft Systems Methodology (Checkland 1972) as a way of addressing changing, ill-defined problem situations.

5. Expression approach

The DRS as described above provides a way to achieve a largely replicable enumeration, being an absolute scale with the logarithmic nature having a tendency to smooth out minor perceptional variations as insignificant figures in the mantissa. However, the DR on its own as a way to express risks and their associated adversities does not address the so-called "magic number" problem.

5.1 Prediction problem

A DR is inherently an expression of an uncertain quantity, with both the likelihood/frequency of occurrence and magnitude of impact being subject to knowability concerns, as highlighted by the work of King (1989) and Gomory (1995).

A particular challenge for much of the fast evolving cyber domain is the poor actuarial detail available for many cases (Bienera, Elinga and Wirfsa 2015) which further distorts the fundamental prediction problem that even where historic data does exist, extrapolation of such data as a forecast of the future is often a questionable approach. This latter concern is amplified in that with ever growing interconnectedness within the cyber domain and in a cyber-physical context, the resultant assemblages are intrinsically complex dynamic systems, which implies a potentially chaotic behaviour to unexpected inputs such as the realisation of an AF. Furthermore, in cases where expert judgement is used, the efficacy of such judgements is also questionable (Tetlock 2006, Gardner 2011, Silver 2013). Engagement with stakeholders across a variety of domains suggested that the 3 Point Estimate[4] approach was generally perceived to be an optimal blend of simplicity and detail.

5.2 Numerical expression

It is preferable to consider estimates of both likelihood/frequency of occurrence and magnitude of impact to be probability distributions, noting that different Adversity Factors and Adversity Classes may have fundamentally differing underlying probability distributions to deal with effects such as LIHP and HILP. The use of a probabilistic approach can partially mitigate, although entirely resolve, the underlying prediction problem. It is therefore proposed that an analogue to the five-number-summary[5], here termed 5-number-prediction (5NP), may be the best way to summarise the expectation (rather than observation) for a likelihood/frequency/magnitude across a variety of types of probability distributions, and, importantly, to not lose the essential outlier characteristics of HILP (e.g. Fat Tails and Black Swans).

As a direct analogue of the five-number-summary, the 5NP consists of:

- Predicted extreme lower value
- Predicted first quartile
- Predicted median
- Predicted third quartile
- Predicted extreme upper value

5.3 Visualisation

The five-number-summary is typically shown as a Box Plot, which encapsulates the elements of the summary diagrammatically with the “whiskers” representing the minimum and maximum, the box the lower and upper quartiles, and the vertical line the median, as illustrated in Figure 2.

Figure 2: A Box Plot

In order to understand a adversity, both the likelihood/ frequency of occurrence and magnitude of impact are needed, yet as we have established that both are probability distributions, we need to extend the one dimensional 5NP / BoxPlot into a second dimension. For this, it is proposed to make an adaptation of the box plot which shows both the uncertainty and the associated adversit: the “10NP” – a matrix of (2 * 5NP).

This adaption may be termed to be a Conquad Plot (derived from the latin *conquadro*: square), as illustrated at Figure 3.

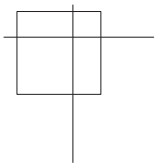


Figure 3: A Conquad lot

[4]Use of a double-triangular distribution to provide an estimated probability distribution derived from predictions of Best, Worst and Middle Cases

[5]A descriptive statistic about a set of observations, showing sample minimum; first quartile; median; third quartile; sample maximum

5.4 Aggregation

There will inevitably be a need to combine adversity estimates for building an AC from AF (Equation 1) and an AS from AC (Equation 2). For a simple "magic number" view of adversity, and assuming that in most cases the AF and AC are likely to be independent factors, a composed DR could be based on unpacking the logarithm, simple addition, and then reforming the logarithm. However, the addition of related adversities and handling of the probabilistic 5NP/10NP approach is more complex, and is the subject of ongoing and future work; what we have at present is an understanding of the features, that will be the basis for establishing a full calculus.

6. Risk expectation

6.1 Context

Many cyber adversity assessments are clouded by practitioners who assume that the magnitude of adversity can be described in a generic manner, whereas in reality most adversities will vary with Entity, Locale, Archetype[6] and Time[7]. It is therefore difficult to perceive a Generic Risk (R_G). The lowest level of meaning analysis is Entity Risk (R_E), which is in practice what is normally expressed, as the preponderance of focus of most cyber management approaches such as the widely adopted methodology from the major international Standards Development Organisations (ISO/IEC 27001:2013).

6.2 Risk eExpression

The problem in handling of extreme distributions like LIHP and HILP has been mentioned already, and is only in part addressed by the use of a probabilistic approach and the 5NP/Conquad Plot. The expression technique proposed is to quote all risks in terms of an Annualised Expectation of Risk (AER), as a terminology to unify and handle adversity in a manner amenable to common interpretation of adversity across differing communities of interest, which:

- is analogous to the concept of Expected Value (EV) from techniques such as Decision Trees and as such is has an explicit artificiality about it that may aid understanding as it requires more active though that the largely meaningless idea of "Risk" – EV is widely used, and understood to be an abstraction that will seldom if ever be the Actual Value; and
- explicitly produces a temporal span – Annularity – which accommodates the LIHP scenario in particular, by factoring in both likelihood and frequency.

6.3 Aggregation

The lowest level of meaning analysis has been stated to be Entity Risk (R_E), which reflects the preponderance of focus of most cyber management approaches as being purely internal. However, the realities of the increasingly interconnected cyber domain and cyber-physical systems means that the abstraction of confining a risk to a single entity is unhelpful, as although a selfish view may seek to only consider the internal factors and ignore externalities, reciprocation of such a view has unfortunate consequences, as the external parties will feel no obligation to share the risks to the focus entity that they may create. Of course, there needs to be consideration of issues such as Joint Risk (R_J), where a combined function is operated (be in a multi-company venture or a multi-national military operation), and a National Risk (R_N) which will be of concern to sovereign state governments, who have to look at the public good impacts from all sources.

The generalised statement of a more holistic approach is provided in **Equation 3**:

$$AER_T = (AER_{AS.E}) + \sum_{1}^{n}(AER_{AS.P}) + \sum_{1}^{n}(AER_{AS.C})$$

[6] A function of the various types of persons engaged in the entity's operation, for instance frequent travellers having a differential risk to those who are static

[7] A function of time, for instance "Y2K" or a major sporting event in which the entity is engaged or is proximate to

where: AER_T – the Total AER either for a single entity and all its externalities, or construct like AER_J (Joint) or AER_N (National)

$AER_{AS.E}$ – the AER from the internal (direct) Adversity Set

$AER_{AS.P}$ – the AER from the Partner(s) (indirect) Adversity Sets

$AER_{AS.C}$ – the AER from Collateral Adversity Sets

As with the case of building an AC from multiple AF, the summation of the preferred probabilistic 5NP approach to AER is complex, and is the subject of ongoing and future work.

6.4 Socialisation

Understanding of combined approach of adversity in the wider community, rather than stovepipes such as Hazard and Threat, has proved to be an ongoing challenge, for despite the obvious attractions of alignment having been identified (e.g. Brostoff and Sasse 2001, and Firesmith 2003). When it comes to practical realisation of scenarios where both main types of adversity are encountered such as National Risk Registers (e.g. Cabinet Office 2012), they are still presented in an isolated manner. It is suggested that adoption of the AER concept as a way of combining the overall risk expectation could become a useful socialisation tool for larger context, analogous to recent popularisation of Micromorts (Howard 1980) and Microlives (Spiegelhalter 2012) for comparing and combining personal risks.

7. Conclusions and further areas for investigation

This paper provides a cross-disciplinary approach to modelling and expressing adversity as part of replicable and scalable methods for the enumeration of risk, and introduces a number of concepts:

- The Adversity Model (AM) as a way to collate differing types of adversity (Hazards and Threats)
- The Adversity Set (AS), consisting of Adversity Classes (AC) and Adversity Factors (AF) identified in the AM
- The Deleterious Result (DR) as an absolute scale for enumeration of impact
- The 5 Number Predication (5NP) as a way to reflect the probabilistic nature of risk, and the associated Conquad Plot which adds uncertainty as a 10 Number Predication (10NP)
- The Annualised Expectation of Risk (AER) as a way to facilitate common interpretation of adversity across differing communities of interest, both internally and in a way to reflect externalities

This work can form the basis of realising a full calculus but further investigation will be required. We suggest that the next steps for this work include consideration of the aggregation of related adversities. It is recommended that there is calibration of probabilistic 5NP/10NP against larger data sets along with a review of the way in which convergence to the mean does or does not occur in such analyses. A further area for work is the production of the Adversity Model and simplification into Adversity Classes is currently a manual process; consideration is required as to the possibility of providing tooling support

References

Ackoff, R.L (1974) Redesigning The Future, Wiley

Aven, T. (2012A), The risk concept— historical and recent development trends, Reliability Engineering & System Safety 99: 33-44.

Aven, T. (2012B), Foundational issues in risk assessment and risk management, Risk Analysis 32.10: 1647-165

Baskerville.R L and Im G P (2005) A longitudinal study of information system threat categories, ACM SIGMIS Volume 36 Issue 4, Fall 68-79

Bienera C, Elinga M, and Wirfsa J H (2015), Insurability of Cyber Risk: An Empirical Analysis, The Geneva Papers 40, 131–158

Brostoff, S and Sasse, M A, (2001) Safe and sound: a safety-critical approach to security, Workshop on New Security Paradigms, ACM 41-50

Bryant, I.R.C. (2012) A Pareto Approach to Software Dependability, I R C Bryant, NATO Research & Technology Organisation Symposium on Information Assurance and Cyber Defence, Germany, September

Bryant, I.R.C. and Watson, T.P. (2014) Replicable and Scalable Adversity Enumeration, NATO Science & Technology Organisation Symposium on Cyber Security Science and Engineering, Estonia, October

Cabinet Office (2012), UK National Risk Register for Civil Emergencies (2012), 17 February

Checkland, P.(1972) Towards a systems -based methodology for real-world problem solving, J.Sys.Eng. 3, 87-116

Cox, Jr, L. A. (2008), Some Limitations of "Risk = Threat × Vulnerability × Consequence" for Risk Analysis of Terrorist Attacks. Risk Analysis, 28: 1749–1761

Everett, H (1957) Relative State Formulation of Quantum Mechanics", Reviews of Modern Physics 29: 454–462

Firesmith, D G (2003) Common Concepts Underlying Safety Security and Survivability Engineering, Software Engineering Institute ADA421683Gardner, D. (2011) Future Babble: Why Expert Predictions Fail and Why We Believe them Anyway, Virgin

Gardoni, P, and Murphy, C (2014) "A scale of risk." Risk Analysis 34.7: 1208-1227.

Gomory, R. (1995) The Known, the Unknown and the Unknowable, Scientific American, June

Howard, R.A. (1980) On making life and death decisions - Societal Risk Assessment: How Safe Is Safe Enough?, General Motors Research Laboratories

Hubbard, D.W, (2009) The Failure of Risk Management, Wiley, April

Hubbard, D.W, (2010) How to Measure Anything, Wiley, May

ISO/IEC 27001 (2013) Information technology -- Security techniques -- Information security management systems – Requirements

ISO/IEC 31000 (2009) Risk management -- Principles and guidelines

Janicb, M and Netjasova, F (2008), A review of research on risk and safety modelling in civil aviation, Journal of Air Transport Management, Volume 14, Issue 4, July 213-220

Judicial College (2013), Guidelines for the Assessment of General Damages in Personal injury cases, 13th Edition

King, J.B (1989) Confronting Chaos, J Bus Ethics 8(1) 39-50

Niederée, R (1992), What do numbers measure? A new approach to fundamental measurement, Mathematical Social Sciences 24:237-276

OECD (2012) The Value Of Statistical Life: A Meta-Analysis, Ecole Nationale de la Statistique et de l'Administration, January

Perrow, C (1984) Normal Accidents: Living with High Risk Technologies, Princeton University Press

Rittel, H W J and Kunz, W (1970) Issues as Elements of Information Systems, Heidelberg

Savage, S.L. (2002) The Flaw of Averages, Harvard Business Review, November

Savage, S.L. (2009) The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty, Wiley, July

Silver, N. (2013) The Signal and the Noise: The Art and Science of Prediction, Penguin

Stevens, S.S (1946) On the Theory of Scales of Measurement, Science Vol. 103, No. 2684, June

Spiegelhalter, D. (2012) Using 'microlives' to communicate the effects of lifetime habits, D BMJ 345, December

Taleb, N.N. (2007) The Black Swan: The Impact of the Highly Improbable, Random House, July

Tetlock, P.E. (2006) Expert Political Judgment: How Good Is It? How Can We Know?, Princeton University Press

On Cyber Dominance in Modern Warfare

Jim Chen and Alan Dinerman
DoD National Defense University, USA
jim.chen@ndu.edu

Abstract: Cyberspace becomes more and more important in modern warfare. It is almost impossible to launch a war without utilizing cyber capabilities in this era. In which way is cyber warfare different from or similar as conventional warfare? What are the unique characteristics of cyber warfare? What roles can cyber play in modern warfare? What are cyber capabilities? How can these capabilities be utilized in deterrence, defensive operations, and offensive manoeuvres? Ultimately, what is cyber dominance? How can cyber dominance be achieved? These are the questions that this paper intends to address. After conducting the literature review, this paper proposes a mechanism in revealing what cyber can do and cannot do in modern warfare. Based on this analysis, it recommends ways of fully utilizing cyber capabilities. This study can help commanders, strategists, and policy-makers to identify, allocate, and make full use of tangible and intangible cyber capabilities in decision-making.

Keywords: cyber capabilities, cyber dominance, cyber warfare, conventional warfare, decision-making

1. Introduction

The term “cyber war” is becoming more prevalent in foreign policy discussions. Increasingly, policy-makers view cyber as an elegant tool to achieve national objectives that can supplant an extensive need for land, sea, and air, and space power. This notion is misnomer because it paints an unrealistic capacity of cyber power to exclusively shape adversaries’ actions. Admiral William McRaven, former Commander of the US Department of Defense’s Special Operations Command, laments that “the enemy’s will, that ultimate center of gravity, remains tied to the ground upon which he sits, upon which he blogs, and to the dirt under his feet” (Freedburg 2013). McRaven continued that “some of the strategists, some of the futurists, want to point to the importance of the social media and the blogosphere and the self-synchronizing organizations - for example, the Twitter-coordinated protests of the Arab Spring - but the fact is geography, terrain, matters.” Russian operations in Georgia and the U.S. operations in Iraq have demonstrated that cyber capacity does not replace the need for land, sea, air, and space capabilities to achieve national objectives. However, cyber power is now an essential component of modern warfare. Success in future warfare will require unity of effort integrating cyber power with traditional power on the land, sea, air, and space domains. To better utilize cyber power and achieve unity of effort, it is important to first understand the similarities and differences between cyber power and conventional compulsory power, the unique characteristics of cyber power, as well as what cyber can do and what it cannot do currently.

Prior to engaging in the aforementioned discussion, it needs to be clarified how the term “cyber power” and the term “cyber dominance” are defined. An operational definition for the term is essential when discussing cyber power in the context of warfare. Strategists and policy-makers now include a vast array of cyber functions under this umbrella. The term “cyber power” frequently encompasses protecting information and communications technologies (ICTs) from cyber attacks, intelligence gathering via networked ICTs, forensics to develop attribution, and/or enhanced military situational awareness / command and control (C2) via net centricity. As it specifically focuses on the offensive operational aspect of cyber power, this paper discusses cyber power in the context of an ability to gather intelligence and execute disruptive offensive effects via networked ICTs. A universal definition for cyber dominance does not exist. Unlike in the land, sea, air, and space, cyber dominance cannot be viewed as complete control of the domain, at least at present. Because of the inter-global connectivity of the cyber domain, because of the fact that the cyber domain is largely compromised of privately owned commercial services, and because of the free-scale nature of cyber results in an ever changing domain, it is hard to expect that any nation will be able to exclusively dominate the cyber domain. Stytz and Banks (2014) provide a more useful definition for cyber dominance. They contend that cyber dominance should be viewed as the ability to control critical elements in cyberspace at a critical time.

This paper examines these concepts. It is organized as follows: In Section 1, an introduction is provided together with the definitions of some essential terms used in the paper. In Section 2, related work is examined. In Section 3, a mechanism is proposed to conduct a comparison between conventional warfare and cyber warfare with

respect to what each can do and what each cannot do in modern warfare. In Section 4, ways of fully taking advantages of cyber capabilities in warfare are discussed. In Section 5, a conclusion is drawn.

2. Related works

Cyber warfare possesses some unique characteristics. Instead of employing conventional weapons such as tanks, warships, warplanes, and missiles, it resorts to ICTs that are comprised of software, hardware, and firmware. As stated in Nye (2010), cyber power "is the ability to obtain preferred outcomes through use of electronically interconnected information resources of the cyber domain". This definition indicates that the means used in cyber warfare are different from those used in conventional warfare but the ends may be the same. Van Houten (2010) shares the same view by maintaining that the purpose of cyber warfare, just like the purpose of conventional warfare, is to use "essentially any act intended to compel an opponent to fulfil our national will". In describing cyber conflicts, Schaap (2009) associates network-based capabilities with "disrupt, deny, degrade, manipulate, or destroy information resident in computer and computer networks, or the computers and networks themselves". Cetron and Davies (2009) observe that "major concern is no longer weapons of mass destruction, but weapons of mass disruption". Andress and Winterfeld (2014) claim that in cyberspace, "the traditional physical boundaries disappear", unlike in conventional warfare where "the two sides operate within the same geographical area".

In order to have a better understanding of cyber warfare and its consequences, one needs to understand cyber capabilities. The works mentioned above have varied focused and are from different perspectives. For example, Schapp (2009) as well as Cetron and Davies (2009) are mainly focused on consequences; while Nye (2010) and Van Houten (2010) are focused on purpose. They are not specifically about cyber capabilities but they are close to the discussion of cyber capabilities. The literature search does not return adequate results on specific description about cyber capabilities, especially the metrics used to compare cyber war capabilities with conventional war capabilities.

In ITU Information Document (2006), a methodology for measuring the capability to counter cybersecurity-related offenses is to use direct indicators, such as technical expertise and clear national cybersecurity policy, and indirect indicators, such as cyber savvy, cyber awareness, and evidence of the dissemination of a culture of cybersecurity.

Rattray and Healey (2010a) hold that cyberspace, as a war-fighting domain, possesses a few key aspects, which are: "logical but physical", "usually used, owned, and controlled predominantly by private sector", "tactically fast but operationally slow", "a domain in which the offense generally dominates the defense", and "fraught with uncertainty". They claim that offensive cyber operations "can be categorized according to a number of factors", which are nature of adversaries, nature of targets, target physicality, integrated with kinetic, scope of effect, intended duration, openness, context, campaign use, initiation responsibility and rationale, initial timing, and initiation attack.

In Rattray (2010b), elements with the cyber environment are compared with other environments. The metrics used are "technological advances", "speed and scope of operations", "control of key features", and "national mobilization".

It is clear that to conduct such a comparison a set of metrics needs to be developed. The next section is focused on creating such a set of metrics.

3. Analysis

As shown above, in Rattray and Healey (2010a) and Rattray (2010b), the factors utilized for the discussion of the cyber domain are nature of targets, target physicality, operation speed, scope of effect, intended duration, etc.

These factors can help us to build the matrices for the discussion of the capabilities of cyber warfare. This paper proposes the use of the following basic set of matrices. Other items can be added into this set if needed. This basic set of matrices consists of who, what, when, where, how, and why. The item "who" is used to inspect the number of people directly involved in conflicts, the number of people directly impacted, and the winners of conflicts. The item "what" is used to examine the targets in conflicts, the cost of conflicts, the characteristics of conflicts, the attribution in conflicts, the rules of engagement, the impression of conflicts, the damage of

conflicts, the deterrence, the dominance, and the result of conflicts. The item "when" is used to scrutinize the preparation time for conflicts, the duration of conflicts, and the time for recovering from the consequences of conflicts. The item "where" is used to inspect the geo-locations of conflicts and the affected areas (or scale) of impact. The item "how" is used to examine the type of strategy used in conflicts. The item "why" scrutinizes the type of purpose for conflicts.

Using these items/parameters, a table can be created to show the differences between conventional warfare and cyber warfare.

Table 1: Conventional warfare versus cyber warfare

	Conventional Warfare	**Cyber Warfare**
Purpose (why)	Gaining political, economic, ideological, social, religious dominance via geo-location dominance for a period of time	Assisting in gaining political, economic, ideological, social, religious dominance; gaining information for competitive advantage
Strategy (how)	Using overt operations and/or covert operations; showing might; little attribution issue	Using overt operations and/or covert operations; attribution issue
Involvement (who)	Some people such as military or paramilitary personnel	Everyone who has a device connected to affected networks
Targets (what)	Humans; mainly tangible objects; directly affecting human life	Mainly intangible items such as information or tangible items such as information systems; may indirectly affecting human life in cyber physical cases
Space (where)	Limited geo-location	Anywhere with respect to geo-location if connected
Duration (when)	A limited period of time	On-going, but one attack is usually within a short period of time
Preparation time (when)	A relatively long period of time	A relatively short period of time
Cost (what)	Expensive	Relatively less expensive
Characteristics (what)	Relatively more transparent	Relatively opaque and in stealth mode
Attribution (what)	Relatively easy to find out	May be hard to find out
Rules of Engagement (what)	Relatively clear	Not clear
Impression (what)	Always severe or brutal; obvious	Less severe if not life & death situation; sometime not felt
Damage (what)	Severe with physical casualty	Severe with information loss
Direct Impact upon (who)	Someone/some businesses	Everyone/every business connected to affected networks
Impact based on (where)	Geo-location	Connection
Deterrence (what)	Obvious and forceful	Limited currently
Dominance (what)	Could be achieved	Hard to be achieved
Result/Gain (what)	Obvious	May not be very clear
Winner (who)	Clear to identify	May be hard to decide
Time for recovering (when)	Relatively long	Relatively short

As shown in this table, in a cyber conflict/war, anyone who has a device may be directly involved or only a few people who has control over a great number of devices, including zombies, may be directly involved. The consequence of a cyber conflict/war may have impact upon everyone connected to the segments being attacked. Under some circumstances, the winner of a cyber conflict is hard to be determined. The targets of a cyber conflict are usually information systems used for all walks of life, or occasionally cyber-physical systems such as industrial control systems. Most cyber attacks are relatively opaque and in stealth mode. Hence, it is difficult to find out who launched the attack. This makes it difficult to apply the rules of engagement, which are also not clear currently. A cyber attack appears to be less severe than a conventional attack if it is not in a life-and-death situation. In most cases, what are damaged are information systems and the information contained within information systems.

In a cyber conflict, the targets are usually information systems and/or information contained within these systems. Should an information system be connected to an industry control system, the ultimate target could be that control system. The cost of a cyber conflict is relatively less expensive than the cost of a conventional

conflict. In most cases, a cyber attack is in a stealth mode, making it difficult to identify real attackers and to apply the rules of engagement, which are also not clear in some cases. Generally speaking, a cyber conflict may give people the false impression that it is not that serious as no lives might be lost. The damage caused by a cyber conflict is severe with information loss or availability of information systems but not severe with physical casualty unless it is a serious cyber-physical attack against an industry control system (ICS) or a supervisory control and data acquisition (SCADA) system. The cyber deterrence is limited currently. The cyber dominance is hard to be achieved. The winner of a cyber conflict is hard to be determined in many cases.

The preparation time for a cyber conflict is relatively short compared with the preparation time for a conventional conflict. So far, a cyber war usually lasts for hours, days, or weeks, but a stealth cyber attack such as advanced persistent threat (APT) may last for months. The recovering time from a cyber conflict is relatively shorter than that from a conventional conflict. The examples can be seen in Richards (2016), which provides a good explanation of the Estonian cyberwar in April-May 2007; in Hollis (2010), which illustrates the cyber war against Georgia in 2008; and in ICS-CERT alert (2016), which describes cyber attacks against Ukrainian critical infrastructure on December 23, 2015. In these cases, cyber wars lasted for relatively short periods of time.

Both the location and the affected areas of a cyber conflict are not restricted by geo-locations. Instead, they can easily be extended to any devices connected to the Internet in any geo-locations.

In most cases, covert operations in virtual environments are used in cyber conflicts, so that attribution is always an issue in these cases. This is how cyber conflicts differ from conventional conflicts; even covert operations are employed in conventional conflicts.

It is interesting to notice that the ultimate purposes of both cyber wars and conventional wars are the same, as both are the tools used in gaining political, economic, ideological, social, and/or religious dominance. Of course, there are some slight differences between the two. The dominance gained via conventional warfare can last for a long period of time while the dominance gained via cyber warfare can only last for a short period of time, at least at present. However, cyber warfare can help to gain critical information for competitive advantage.

Based on this comparison/analysis, a sketch of cyber warfare can be drawn. It is a tool that can be used in gaining political, economic, ideological, social, and/or religious dominance. It is good at helping to gain critical information for competitive advantage, as covert operations in virtual environments are utilized in most cases. In cyber warfare, one can get to targets within a short period of time in a wide scale, but the long-lasting effects of cyber campaigns are limited or restricted. However, it might generate some unexpected effects that are hard to be achieved via conventional warfare, as shown in the cyber-physical environments, such as ICS/SCADA systems or Internet of Things. Deterrence can be achieved through this kind of surprise effects. In addition, launching a cyber war is not as expensive as launching a conventional war.

It takes quite some time and costs a lot more money to launch a conventional war but such a war can generate long-lasting effects. If some unique cyber capabilities, such as intelligence collection, stealth manoeuvres, and surprise effect, are included in a joint operation, the military capabilities will be greatly increased. Evidently, cyber capabilities and conventional warfare capabilities are complementary to each other. An integration of both capabilities will yield even greater capabilities.

The next section is focused on the discussion about how to fully take advantages of the strengths of cyber capabilities.

4. Discussion

The analysis in the previous section shows both cyber capabilities and conventional capabilities have their unique characteristics and in some areas they are complementary. If both capabilities are integrated together, stronger military capabilities can be generated.

As discussed previously, the notion of cyber war is really a misnomer. Nations will not execute war via cyber exclusively as the capabilities that cyber possesses at present do not match exactly the capabilities of conventional warfare. Terrain will continue to matter. Land, sea, air, and space power will remain essential components of compulsory power. However, in future conflicts, cyber power will become increasingly

important. Nation states will synchronize cyber capabilities with traditional land, sea, air, and space capabilities in order to achieve objectives. Cyber power has unique characteristics that allow a great operational advantage when effectively synchronized with traditional land, sea, air, and space power. This operational advantage is most manifest in two areas: (1) an ability to achieve an asymmetry that offsets numerical advantage; and (2) an ability to offset terrain in order to execute deep strategic strike.

4.1 Achieving asymmetry

Typically the concept of asymmetrical warfare is mostly associated with insurgents, as opposed to, nation states. However, this paradigm is not totally comprehensive. The RAND cooperation defines asymmetrical warfare as "conflicts between nations or groups that have disparate military capabilities and strategies" (RAND, 2016). While many Western nations enjoy military superiority, their military equipment and personnel volume pale in comparison to some nations. In essence, Western nations have a disparate military capability in terms of quantity. Integrating cyber operations with operations in the land, seas, air, and space provides Western nations an asymmetry that offsets its numerical disadvantage.

Observing the United States success with net-centric operations during Desert Storm, other nations have begun retooling their doctrines and command and control (C2) methodologies to take advantage of network integrated platforms. Some nations have aggressively modernized their command, control, communications, computers, intelligence, surveillance, and reconnaissance programs and adjusted doctrine to incorporate cyber capabilities. While this type of modernization can enhance military capacity, it simultaneously introduces vulnerabilities that are exploitable via cyber operations. Cyber operations, which disrupt networked information operations, greatly diminish adversarial numerical advantages and accentuate capacities in the land, sea, air, and space domains.

4.2 Enabling deep strike

A nation's desire to execute deep strike that disrupts industrial and critical infrastructure is not new to the 21st century. In the 19th century, military commanders unleashed horse-mounted cavalry to quickly get behind enemy lines and destroy food storages, lines of communication, or arms factories. The 20th century ushered in the era of air power, enabling a tremendous reduction in the time-space calculus needed to strike at critical infrastructure. During World War II, strategic bombing of cities and factories emerged as a seminal American operational strategy. Whereas horse mounted cavalry would have needed days to manoeuvre from France to Germany, air craft could execute strategic bombing in hours.

Cyber further reduces the time-space calculus to net speed. This is not, however, cyber power's greatest contribution to deep strike. In both the cavalry and aircraft employment, holding terrain was a significantly limiting factor. Both horse and aircraft were limited by the geography held. Cyber power allows nations to cripple critical infrastructure and line of communications at net-speed from thousands of miles away. In 21st century warfare, cyber power allows nations to conduct deep strike shaping operations without first securing geographical terrain.

These two examples clearly show how cyber capabilities can be employed to support conventional military capabilities in offensive manoeuvres. The integrated capabilities are more powerful. It helps to achieve military dominance, not necessarily cyber dominance. In other words, cyber dominance can only be achieved via its integration into conventional military capabilities. When they are in synchronization, new powerful capabilities can be generated. This also means that the joint military concept needs to include cyber capabilities.

5. Conclusion

It is shown in this paper that cyber warfare possesses some unique characteristics; as these characteristics do not exactly match those of conventional warfare, a cyber war that is executed exclusively is hard to imagine. However, cyber capabilities, if integrated appropriately into conventional warfare, can serve as force multipliers as unique cyber capabilities such as intelligence collection, stealth manoeuvres, and surprise effect are complementary to the existing military capabilities. At least at present, exclusive cyber dominance is also hard to imagine, but successful use of well-integrated joint military capabilities in the five domains (land, sea, air, space, and cyber) can eventually lead to military dominance.

A good understanding of these points can helps commanders, strategists, and policy makers to identify, allocate, and make good use of unique cyber capabilities in decision-making. Integrated and joint capabilities can serve as force multipliers.

References

Andress, J. & Winterfeld, S. (2014) *Cyber Warfare: Techniques, Tactics, and Tools for Security Practitioners*. 2nd Edition, Amsterdam: Syngress, an imprint of Elsevier.

Cetron, M. & Davies, O. (2009) "Ten Critical Trends for Cyber Security". *The Futurist*, Volume 43, Issue 5, Pages 40-49.

Freedburg, S. (2013) "People, Cyber, and Dirt: Army and SOCOM's Strategic Landpower", Breaking Defense. Retrieved from http://breakingdefense.com/2013/10/people-cyber-dirt-army-socoms-strategic-landpower.

Hollis, D. (2011) "Cyberwar Case Study: Georgia 2008", Small Wars Journal. Retrieved from http://smallwarsjournal.com/blog/journal/docs-temp/639-hollis.pdf.

ICS-CERT, U.S. Department of Homeland Security. (2016) "Cyber-Attack Against Ukrainian Critical Infrastructure", Alert (IR-ALERT-H-16-056-01). Retrieved from https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01.

Morel, B. (2006) "A Methodology for Measuring the Capability to Counter Cybersecurity-Related Offenses", ITU Information Document. Retrieved from https://www.itu.int/osg/csd/cybersecurity/2006/morel-paper-15-may-2006.pdf.

Nye, J. (2010) "Cyber Power", Belfer Center for Science and International Affairs, Harvard Kennedy School. Retrieved from http://belfercenter.hks.harvard.edu/files/cyber-power.pdf.

RAND Corporation. (2016) "Asymmetric Warfare". Retrieved from http://www.rand.org/topics/asymmetric-warfare.html.

Rattray, G. & Healey, J. (2010a) "Categorizing and Understanding Offensive Cyber Capabilities and Their Use", *Proceedings of a Workshop on Deterring CyberAttacks: Informing Strategies and Developing Options for U.S. Policy*, Pages 77-97, The National Academies.

Rattray, G. (2010b) "An Environmental Approach to Understanding Cyberpower", *Cyberpower and National Security*, Pages 253-274, Kramer, F. et al. (Eds.), National Defense University Press and Potomac Books, Inc.

Richards, J. (2016) "Denial-of-Service: The Estonian Cyberwar and Its Implications for U.S. National Security", the International Affairs Review, the Elliott School of International Affairs at George Washington University. Retrieved from http://www.iar-gwu.org/node/65.

Schaap, A. (2009) "Cyber Warfare Operations: Development and Use under International Law", *Air Force Law Review*, Volume 64, Pages 121-174.

Stytz, M. & Banks, S. (2014) "Toward attaining cyber dominance", *Strategic Studies Quarterly*, Spring 2014, Pages.55-87.

Van Houten, V. (2010) "An Overview of the Cyber Warfare, Exploitation & Information Dominance (CWEID) Lab". Retrieved from http://info.publicintelligence.net/cyberwarfarebrief.pdf.

Military Strategy as a Guide for Cybersecurity

Allen Church
The MITRE Corporation, McLean, USA
achurch@mitre.org

Abstract: Current practices aimed at addressing the vulnerability of modern information systems to cyber-attack have been described as "whack-a-mole." The term refers to a "situation in which attempts to solve a problem are piecemeal or superficial, resulting only in temporary or minor improvement" (Oxford English Dictionary). In many cases, information security professionals engage in seemingly endless tactical efforts to: 1) fix software weaknesses that may be subject to cyber exploitation; 2) monitor all network traffic at the packet-level for evidence of malware code; and 3) apply a growing number of security controls and tools to all the computational components of their system. As an example of the latter, the current list of controls published by the U.S. National Institute of Standards has grown to about one thousand. This paper will argue for adopting a more strategic approach to cybersecurity, and will use examples drawn from both historical and modern writings on military strategy to buttress its position. Historical sources will include von Clausewitz (*On War*, ca 1831) and Condell and Zabecki's (2001) translation of *Truppenführing*. Also included is a brief mention of the ideas of U.S. Air Force Colonel John Boyd, who is best known for crafting the Observe, Orient, Decide, and Act (OODA) loop that has been applied to business as well as military strategy. The OODA loop provided a starting point for a set of strategic cyber security elements that include: build layers of defense, enhance situational awareness of defensive status, manage settings and software, and build operational teamwork and morale. The concept of time provided an overarching principle for tying these different elements together. The goal of this article is to create a strategic model for cyber security that is based on enduring principles and described by a simple unifying concept. It is anticipated that this model will be improved over time and extended to specialized areas as needed. An example of an element in the current framework that will need to be improved became evident following comparisons against the strategic principles handed down by von Clausewitz, in particular.

Keywords: strategic cybersecurity, military strategy, defensive principles, situational awareness

1. Introduction

Recent history has shown that public and private sector information systems have been the subject of an increasing number of cyber-attacks, resulting in the theft of customer financial data as well as the loss of significant amounts of personally identifiable and protected-health information. In an attempt to identify some of the factors that have contributed to the susceptibility of information systems to such attacks, the current state of cybersecurity practices will be briefly reviewed and analysed at a high-level. In the following section, a simplified strategic framework for guiding cybersecurity defense will be presented. This will be followed by sections that briefly describe each of the framework elements and identify supporting material drawn from historical sources on military strategy and tactics. The primary sources for strategy and tactics were von Clausewitz (1984) and *Truppenführing* (Condell and Zabecki, 2001) respectively.

The intent of this article is to: 1) paint a practitioner's picture of current approaches for achieving cyber security; 2) identify problem areas revealed by that picture; 3) craft a strategic framework to simplify the pursuit of cyber security; and 4) compare that framework against historical sources on military strategy and tactics in order to refine it. The results showed that the first two elements of the framework, namely: building layers of defense and hiding their purpose; and enhancing situational awareness, were well-supported by previous military strategy. The third element (managing settings and software) received little historical support and will need to be replaced in a future version of the framework. Finally, the fourth element (originally termed "Improve Operational Efficiency") was refined to emphasize the importance of teamwork and esprit de corps in accordance with insights provided long ago by von Clausewitz.

[*Guide to the reader*: For quotations from the work of von Clausewitz, page references will be made to the Kindle e-book Location and abbreviated as Loc. For quotations from Condell and Zabecki, page numbers will be used and the source will be identified as *Truppenführing*.]

2. Current cybersecurity practices

During the course of advising several organizations regarding the cybersecurity status of their information systems, a number of general characteristics were evident:

- Information security is managed as a compliance activity in which a list of security controls are compared against those that are present in the information system being assessed.

- Information security controls and processes are described in guidance (SP-800-53r4, 2013) provided by the U.S. National Institute of Standards and Technology (NIST).
- The number of individual security controls has grown substantially over time as NIST has identified additional characteristics within an information system that may be subject to exploit.
- While the guidance was developed as a comprehensive source of controls for information systems as a whole, the controls are applied and audited at the level of individual system components.
- In an enterprise with a large number of system components, each of which may be subject to a large number of controls, compliance audits can require the review of tens or hundreds of thousands of individual controls.
- One consequence of this large number is that individual controls need to be checked quickly to meet compliance schedules, and thus the quality of each assessment becomes subordinate to the speed with which it can be performed.
- Another consequence is that information security staff may become overwhelmed by the continuous tactical tempo of the work, resulting in psychological "burn-out" and the loss of valuable staff due to increased turnover. Moreover, such turnover aggravates the situation by increasing the workload of remaining staff.
- Staff burn-out and turnover are also factors that affect security operations personnel, due to the demands imposed by continuous monitoring of large amounts of network traffic (e.g., at the packet level).

3. High-Level analysis

Information security controls were initially developed to protect individual systems that supported large applications. Each application needed to have a full suite of application security controls, and the underlying platform had to be defended with network and operating system security controls. With the rise of local area networks, enterprise information systems grew from dedicated islands of terminals and mainframes to shared networks with personal computers (PCs) and client-server applications. At this stage, there was a jump in computing power, but it came with a jump in the number of computers that needed some measure of security management. As enterprises connected to the Internet and adopted easy-to-use web technology, the number of applications grew quickly along with the number of web servers. As networks became more complex, additional technology was added to enable all users to access the Internet and internal computing resources, and to extend connectivity to portable wireless clients. While all of these capabilities have enhanced the ability of users to obtain and process data, they have also added to the number of computers that need to be managed and monitored. At present, it is evident that the management of all these separate systems and applications within an enterprise has become challenging and expensive. In addition to the burden of managing all these smart devices, network traffic levels have grown to the point where “natural human capacities are becoming increasingly mismatched to data volumes, processing capabilities, and required decision speeds” (CV-2025, 2012).

When this large number of hardware and software components and associated inter-system dependencies are considered from a control perspective, it no longer seems to make sense to simply increase the number of system-level controls and monitoring tools that are applied to manage them. We contend that this will lead to a further increase in the complexity of current information systems and a corresponding decrease in our ability to manage them. If one steps back and constructs a high-level view of systems that move large volumes of network traffic across a diverse array of interconnected hardware and software components, the resulting picture is one of hyper-complexity. At face value, this picture suggests that many enterprise systems are currently bordering on a state of being out of control with respect to their management. In support of this position, the facts show that the number and severity of cyber-attacks continue to steadily increase, as does the cost and difficulty of managing information security.

4. Strategic framework

The large number of hardware and software components in enterprise information systems could be limiting our ability to manage them. Moreover, tactical approaches to cybersecurity could aggravate the situation, and result in less ability to resist cyber-attacks. Since current efforts to counter cybersecurity attacks have been largely unsuccessful, we believe that a more strategic approach is needed. Recognizing that cybersecurity depends on the ability to defend information systems against attack, we will use established principles drawn

from historical sources of military strategy and tactics to support our position. Acknowledging in advance that no approach is perfect, we have identified a framework of cybersecurity principles that are relatively easy to understand and remember, while also being strategic in the sense of providing long-term value. Each of the following principles is contained in the proposed framework:

- *Time* (i.e., for the defender) will serve as an overarching performance metric;
- Build Layers of Defense and Hide Their Purpose;
- Enhance Situational Awareness;
- Manage Settings and Software; and
- Build Operational Teamwork and Morale.

This strategic framework for cybersecurity is presented in Figure 1, where the concept of *Time* is illustrated as playing a central role that guides the development of an integrated defense.

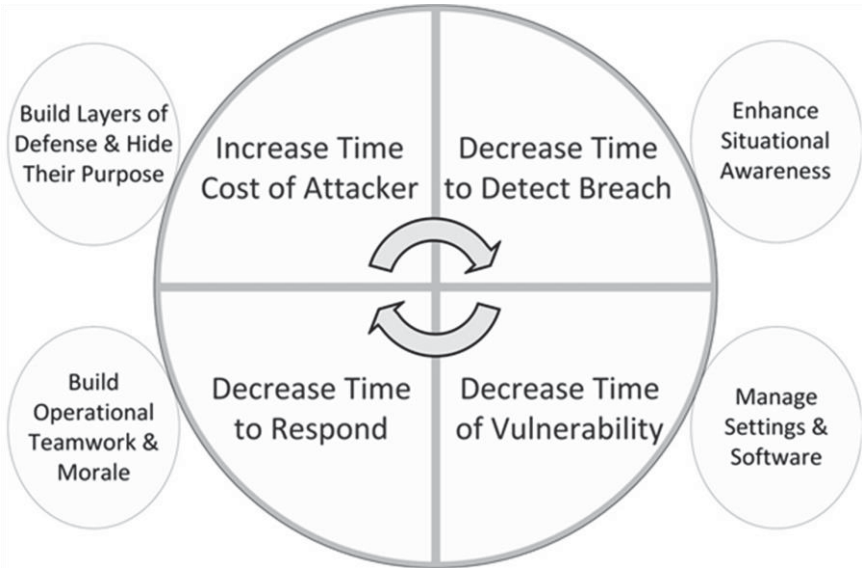


Figure 1: Strategic framework for cybersecurity

5. Time

As indicated above, time is a unifying concept in our approach. A critical role of time was identified by Col. John Boyd with respect to factors that support effective military command and control. He argued that commanders and subordinates should strive to develop a shared set of implicit understandings to (Boyd, 1987):

- Diminish their friction and reduce time [which permits them to]:
- Exploit variety/rapidity while maintaining harmony/initiative [which permits them to]:
- Get inside adversary's Observe-Orient-Decide-Act (OODA) loop, thereby:
- Magnify adversary's friction and stretch-out his time.

Boyd's OODA loop was a concept consisting of four steps that a fighter pilot would need to repeatedly execute in an aerial combat setting. Boyd argued that a pilot would steadily gain a tactical advantage over an adversary, if he/she could cycle through those four steps more quickly than the opposing pilot. Nearly two hundred years ago, von Clausewitz (1984) identified a related capability that had both tactical and strategic value:

> *The aspect of war that has always attracted the greatest attention is the engagement. Because time and space are important elements of the engagement, and were particularly significant in the days when the cavalry attack was the decisive factor, the idea of a rapid and accurate decision was first based on an evaluation of time and space, and consequently received a name which refers to visual estimates only. Many theorists of war have employed the term in that limited sense. But soon it was also used of any sound decision taken in the midst of action—such as recognizing the right point to attack, etc. Coup d'oeil therefore refers not alone to the physical but, more commonly, to the inward eye. The expression, like the quality itself, has certainly always been more applicable*

> *to tactics, but it must also have its place in strategy, since here as well quick decisions are often needed. (Loc. 2133)*

The ability to quickly understand the intent of an action taking place in time and space, and to decisively counter it, were recognized as vital traits for a defensive military commander. The special importance of time for defenders was explicitly recognized by von Clausewitz:

> *"Wearing down the enemy in a conflict means using the duration of the war to bring about a gradual exhaustion of his physical and moral resistance. (Loc. 1940);*
>
> *"...the time that passes is lost to the aggressor. Time lost is always a disadvantage that is bound in some way to weaken him who loses it." (Loc. 7229);*
>
> *"Meanwhile the defender is gaining time—which is what he needs most." (Loc. 7341); and*
>
> *"On the other hand, this type of river defense can often gain considerable time—and time, after all, is what the defender is most likely to need." (Loc. 8359).*

Time was also stressed for the defensive form of warfare in *Truppenführung*, a guide written for German military commanders prior to World War II.

> *"The timely delay of the enemy approach will result in additional freedom of action" (p. 120);*
>
> *"The defensive between lines of resistance must delay the enemy and buy time for the preparation of the next line of resistance." (p. 132); and*
>
> *"The movements of the enemy should be harassed and delayed far to the front of the line of resistance..." (p. 134).*

As these examples point out, an effective defense is one that grants more time for defenders to counter an attack, and one way to accomplish this is by imposing delays (i.e., a loss of time) on an attacking force. This simple principle of optimizing defensive time provides a strategic approach that is an integral part of all the principles that follow. Time serves as a common ruler for gauging their effectiveness.

5.1 Build layers of defense and hide their purpose

A layered defense for an information system relies on the use of multiple types of security controls that are arrayed to create defensive redundancy. By carefully planning how different types of security controls should be positioned, an interlocking set of logical barriers can be created that retards the progress of cyber-attacks. However, it should be recognized that even multiple layers of defense are no guarantee that an information system will be immune to all forms of attack. Nonetheless, a well-planned and executed array of information security controls will greatly diminish the number of ways that a system can be compromised. This will typically slow the progress of an attacker and provide defenders with more time to detect and counter an attack.

The value of layered defensive fortifications was emphasized by von Clausewitz (1984) as an important method for strengthening a defensive position. He described the result as consisting of a combination of reinforced protective structures, trenches and obstacles that served to slow the advance of attackers.

> *The defender waits for the attack in position, having chosen a suitable area and prepared it; which means he has carefully reconnoitered it, erected solid defenses at some of the most important points, established and opened communications, sited his batteries, fortified some villages, selected covered assembly areas, and so forth. The strength of his front, access to which is barred by one or more parallel trenches or other obstacles or by dominant strong points... (Loc. 7395)*

In addition to these physical layers, von Clausewitz (1984) also advised that troops be arrayed in depth:

> *"He holds his position in depth, for at every level, from division to battalion, his order of battle has reserves for unforeseen events and to renew the action." (Loc. 7400).*

The advice contained in *Truppenführing* regarding defensive preparations is very similar to that provided by von Clausewitz:

> *"The main battle area must be organized in depth. Its purpose is to … facilitate the continuation of the defense — even when the attacker has penetrated the main battle area." (p. 122); and*

> *"A well-constructed main battle area normally consists of a chain of mutually supporting positions with obstacles, trenches and individual firing positions. The positions should be distributed irregularly and in depth, and are established in the sequence of their importance." (p. 122).*

While a defense should consist of multiple interlocking layers, both sources also stressed the importance of disguising the type and strength of such fortifications (von Clausewitz and *Truppenführing*, respectively):

> *"In our opinion, a defensive position approaches the ideal the more its strength is masked, and the more it lends itself to taking the enemy by surprise in the course of the action. One always attempts to deceive the enemy as to the true numerical strength of one's fighting forces and their true direction. By the same token, then, one should not let him see how one intends to take advantage of the terrain" (Loc. 7767); and*
>
> *"The enemy should be deceived for as long as possible as to the exact location of the main line of resistance. The terrain, therefore, must be analyzed from the standpoint of enemy observation and the defensive positions adapted accordingly." (p. 123).*

Deception is rarely considered when setting up information security controls, and this leads naturally to the next section on situational awareness. When the nature of security controls can be hidden or disguised, an attacker will have less situational awareness about the type of controls being used, and similarly, less understanding of how they have been logically arrayed.

5.2 Enhance situational awareness

Situational awareness refers to the state of having an accurate understanding of the current conditions that exist in a given environment. In the present case, this environment is frequently limited to the security settings and controls that protect an information system. Situational awareness of information system security should include an understanding of: a) the threats that exist outside the system; b) the integrity and capabilities of its defense; and c) the flow of data moving across the internal network. Awareness of existing cyber threats can be used to tune system defensive measures to resist or counter such threats (e.g., detecting the signature of specific malware). Awareness of the integrity of system security requires monitoring the software and firmware settings that reside on numerous types of devices, and then making appropriate updates to those settings and programs. The latter topic will be addressed in the next section that deals with settings and software management. In the case of internal data flows, we maintain that enterprise situational awareness is incomplete, if the flow of high-value data is not carefully monitored and understood. The reason for this position is that the primary goal of most cyber-attacks is to gain access to high-value data (e.g., financial, personal, proprietary or sensitive operational information). There are abundant examples showing that the exfiltration of data via cyber compromise has gone undetected for months and sometimes years. This point is consistent with the claim made above that many information systems appear to be out of control with respect to their management. Without knowing who is accessing data, what that data is about, when it was accessed and where it went, situational awareness about the presence of an active cyber-attack is limited or non-existent.

The importance of situational awareness for defenders was made very clear by von Clausewitz in the following excerpts:

> *"There is still another factor that can bring military action to a standstill: imperfect knowledge of the situation. The only situation a commander can know fully is his own; his opponent's he can know only from unreliable intelligence." (Loc. 1749);*
>
> *"We conclude that an accurate and penetrating understanding is a more useful and essential asset for the commander than any gift for cunning..." (Loc. 4010);*
>
> *"Above all, the defender must seek to keep the enemy under observation.." (Loc. 7738); and*
>
> *"It is in the interest of the defender, even more than of the attacker, to command an unimpeded view, partly because he is normally the weaker of the two, and partly because the natural advantages of his position lead him to develop his plans later than the attacker." (Loc. 8656).*

This same view was expressed in *Truppenführing* as shown in the following:

> *"Reconnaissance (Aufklärung) should produce a picture of the enemy situation as rapidly, completely, and reliably as possible." (p. 39);*

> *"Good ground reconnaissance also contributes to good security. Conversely, the actions of a security unit produce a certain amount of reconnaissance. Reconnaissance and security on the ground complement one another and cannot be separated. (p. 40);*
>
> *Different methods of reconnaissance supplement one another. The shortcomings of one method are compensated for by the strengths of the others. (p. 42);*
>
> *Since the defensive does not have the advantage of the initiative, it requires the earliest possible contact with the enemy using all means of reconnaissance to determine the direction of his advance and the composition and strength of his forces. (p.119);*
>
> *All arms must accurately monitor the enemy's situation. (p. 123); and*
>
> *The defender must derive the enemy's attack plan as early as possible. He monitors the reconnaissance and intelligence-gathering systems of the attacker, he monitors the enemy radio nets, and he tries to determine the strength and deployment of the approaching forces. (p. 124).*

As these historical views of military strategy and tactics reveal, situational awareness is especially important for defenders. One aspect of situational awareness is an understanding of the status of defensive controls and this topic is discussed next.

5.3 Manage settings and software

A major part of managing IT security involves: a) applying appropriate configuration settings to hardware and software assets; and b) patching software that contains known vulnerabilities. This process depends on maintaining situational awareness regarding the version and configuration settings of hardware and software components in the enterprise, and responding in a way that resolves weaknesses and vulnerabilities of components without breaking system functionality. As currently practiced, this security management activity is often performed without regard to the location or criticality of an asset, and thus, frequently consumes large amount of organizational resources. One of these resources is time.

The resources required to manage all these assets can be reduced by taking advantage of defense-in-depth layers to shield some of the assets from being directly accessible on the network. The use of application-layer gateways to create enclaves is one way to create a fortified defensive position in an information system. Although the following advice from von Clausewitz is not exactly analogous, it speaks to the need for correct analysis, careful planning and prioritization of resources.

> *"Relative superiority, that is, the skillful concentration of superior strength at the decisive point, is much more frequently based on the correct appraisal of this decisive point, on suitable planning from the start; which leads to appropriate disposition of the forces, and on the resolution needed to sacrifice nonessentials for the sake of essentials..." (Loc. 3899).*

The point of view expressed in *Truppenführing* is more tactical in nature, as evidenced by the emphasis on continuous monitoring.

> *"Continuous reconnaissance is the essential element of security." (p. 57).*

However, there is another perspective provided in Appendix E of Condell and Zabecki (2001) that was not part of the original *Truppenführing.* This observation was made by a group of high-ranking German officers that conducted an analysis in 1952 of the U.S. Army Field Service Regulations. They expressed a concern about "security requirements" assuming precedence over the mission needs. This concern is similar in some respects to the point made above about the resource costs of following security rules without regard to asset priority.

> *"The attempt to find a solution for every single situation, which may confront the lower echelons, occasionally results in a cut-and-dried "recipe" which is far more detailed than needed.";*
>
> *"The same applies to "security requirements," which are often exaggerated and frequently cause adherence to a plan which takes priority over the accomplishment of the mission." (p. 283).*

5.4 Build operational teamwork and morale

In this final section on taking a strategic approach to cybersecurity, the factors that contribute to reducing the time required to respond to an attack will be briefly addressed. As shown in Figure 1, two factors are called out for improving the performance of cyber security operations. One factor refers to the ability of a team to work

together to efficiently achieve a shared goal, while the other points to development of a state of personal and team pride. Although technical solutions can supplement achievement of the former factor, the latter factor depends more on the positive psychological state that results from adoption of team values, shared goals and a unity of purpose. Both characteristics will benefit from the early detection of a successful cyber-attack. As previously discussed, situational awareness of the presence of an attacker within an information system will need to be significantly improved to realize its full benefits. On the other hand, the development of high-performance operational teams and improved team morale can be readily achieved through the use of existing training materials and curricula. The benefits of improved teamwork and morale can be further amplified by using existing technology to automate the performance of commonly-executed routines. As mentioned in Section 2, many cyber security professionals are overworked, and thus, every effort should be made to reduce the level of routine demands being placed on these personnel.

Since a heightened state of vigilance is chronically required from cybersecurity operations staff, and negative feedback from upper management is often the rule (e.g., following a cyber compromise), the morale and effectiveness of an operational team will tend to become worn down. As von Clausewitz points out, the value of developing and maintaining team morale (esprit de corps) should not be underestimated.

> *"it would be a serious mistake to underrate professional pride (esprit de corps) as something that may and must be present in an army to greater or lesser degree. Professional pride is the bond between the various natural forces that activate the military virtues; in the context of this professional pride they crystallize more readily." (Loc. 3706); and*
>
> *"Military spirit, then, is one of the most important moral elements in war. Where this element is absent, it must either be replaced by one of the others, such as the commander's superior ability or popular enthusiasm, or else the results will fall short of the efforts expended." (Loc. 3735).*

6. Conclusion

This article has provided a review and analysis of cybersecurity practices and the enterprise information systems that they need to protect. Concerns were raised about the level of complexity resident in many of these systems, how this complexity may challenge attempts to secure such systems, and the likelihood that further tactical cybersecurity efforts will increase system complexity and diminish the ability to effectively manage system cybersecurity. A strategic cybersecurity framework was identified in which time provided a consistent metric for evaluating the strategic principles associated with system defense. Historical sources of military strategy and tactics were examined and compared to the elements contained in the cybersecurity framework. The first two elements (i.e., building layers of defense whose purpose is concealed and improving situational awareness) were strongly supported by historical sources. In contrast, the third element, namely the management of settings and software, found little support with respect to military strategy, and will need to be reconsidered. The last framework element (originally focused on operational efficiency) was revised following the initial draft of this paper to emphasize the importance of teamwork and morale that von Clausewitz identified in the early nineteenth century.

Acknowledgements

The views expressed in this article are based solely on the opinions and research of the author, and do not in any way reflect the views of the MITRE Corporation. The author would also like to acknowledge the excellent comments provided by the individual who performed a double-blind review of the original manuscript.

References

Boyd, John R. (1987) “Organic Design for Command and Control,” May

Condell, B. and Zabecki, D.T. (2001) ***On the German Art of War: Truppenführung***. Boulder, CO: Lynne Rienner Publishers.

(CV-2025) Cyber Vision 2025, United States Air Force Cyberspace Science & Technology Vision 2012-2025, (2012) U.S. Air Force, 1 September.

Oxford English Dictionary, http://www.oxforddictionaries.com/us/definition/english/whack-a-mole.

(SP-800-53r4) Security and Privacy Controls for Federal Information Systems and Organizations (2013) National Institute of Standards and Technology, April.

Von Clausewitz, C. (1984) ***On War***, ed., transl. M. Howard and P. Paret. Princeton, NJ: Princeton University Press.

Applications of Identity Based Cryptography and Sticky Policies With Electronic Identity Cards

Paul Crocker and João Silveira
Universidade da Beira Interior, Portugal
crocker@di.ubi.pt
m4384@ubi.pt

Abstract: This article will describe the implementation of a system for privacy and confidentiality for files and messages using Identity based cryptography in conjunction with Electronic Identity Cards that have as a common characteristic strong government backed authentication mechanisms. The system enables the files to be encrypted for multiple identities, or recipients, and enables privacy and security policies to be associated to the file. These policies are structured Extensible Markup Language files and permit a range of policies based on a users role or access level, device and network information and even time intervals. No prior knowledge or sharing of the recipients public key is necessary as the encryption key is based on the identity of the user derived from the users electronic Identity Card. Electronic Identity cards are becoming standard in many European countries and can be used for proving the holders identity, physically and electronically, and even for creating digital signatures, however standard encryption services are not usually included. In the system proposed here the creation and distribution of private keys is the role of a trusted third party, which the user can delegate to any organization of their choice. Authentication at this key centre is crucial for the security of the system and relies on the strong two-factor authentication of electronic identity cards. The article will describe the algorithms used for encrypting files and messages for multiple users the overall cryptographic system and the formats used for the encrypted files that incorporates a cryptographic hash as well as the XML file policy. The final system is implemented as a C# library for the various algorithms and methods which contains a wrapper for a well-known pairing based library written in standard C. Our library can be used as a plugin for any number of applications; in particular the applications implemented and described in this article are namely a small tool for creating policies, a standard *Desktop* application and also a Cloud service for cloud storage system that can enable the enforcing of the file policies. The proposed systems as well as being innovative provides a secure system for file confidentiality and privacy as well as being transparent and easy to use by the end users of the system.

Keywords: confidentiality, electronic identity, identity based cryptography

1. Introduction

Electronic identity (E-Id) cards permit the holder to prove his or her identity, physically or electronically using a diverse range of mechanisms. This type of card commonly contains human and machine readable information including name address, identity numbers, photograph and often biometric information such fingerprint templates. At a national level these types of cards are based on sovereign state governmental systems that include many levels of public government departments and national Public Key Infrastructure (PKI) for managing the system.

The Portuguese E-Id card or Cartão de Cidadão (CC), which is the E-Id used in this article, was rolled out in 2007 with the main objective of merging various identification documents into a single more secure electronic smart card that as well as offering the tradition physical means of identification also enables greater electronic interaction with the governmental services and provides a legal and electronic infrastructure for validating electronic documents.

Although E-ID cards are highly secure Smartcard containing a pair of (usually RSA) keys for authenticating the identity of the citizen and providing a qualified electronic digital signature they do not in general provide a means of encrypting and decrypting documents using the governments' public key infrastructure. However it's possible to design solutions that work around this problem that make use of characteristics that are present on the card, such as serial-id, personnel id number etc. Using Identity Based Cryptography (IBC) it is possible to use a combination of some of these characteristics present on the E-ID cards to derive a public key, permitting file encryption without the necessity of exchanging certificates or public keys and even without the existence of public key infrastructure for the management of keys.

The challenge is to develop a solution that grantees the privacy and confidentiality of the encrypted data whilst simultaneously giving the end user a high degree of confidence in the solution that is also flexible and simple to use. In this article we show how this may be achieved and also provide new language wrappers for an existing low level elliptic curve pairing library.

This article is divided into six sections. In section 2 related work will be presented. In section 3 an introduction to identity based cryptography is given and the cryptographic schemes used detailed. Details of the implementation of the schemes are given in section 4. The architecture and applications developed in this work are presented in section 5. Finally, in section 6 the conclusions and future work will be discussed.

2. Related work

Identity based Cryptography was proposed in 1984 by Shamir (1984), however Shamir only presented a solution for a signature scheme, leaving open the question of finding a scheme for identify based encryption. Even though various schemes were subsequently proposed none were completely satisfactory until 2001 when Boneh and Franklin (Boneh 2001) presented a completely functional scheme based on bilinear pairings over elliptic curves and finite groups. Cocks (2001) also presented a solution for identity based encryption, based on quadratic residues rather than bilinear pairings. However this scheme has as its main disadvantage the fact that it produces long cypher texts, for instance the encryption of a 128 bit message using a modulus of 1024 bits produces a cypher text of 32678 bytes compared to only 36 bytes for the pairing based scheme. Boneh (2007) also presented a scheme based on quadratic residues, similar to Cocks scheme and with the same security properties that solved the problem of long cypher texts. This scheme results in cypher texts of 145 bytes for messages of 128 bits and using a modulus of 1024 bits.

In order to send a message to *n* receivers the Boneh-Franklin scheme needs to encrypt the message *n* times and thus perform *n* pairings. On the other hand Baek, Safavi-Naini e Susilo (Baek 2005) presented a scheme based on pairings over elliptic curves in 2005 that enables the encryption of messages for multiple receivers identities that only requires one pairing to encrypt a single message for *n* receivers

Examples of some real world cryptographic applications using IBC can be found in the email system developed by *Voltage Security* (www.voltage.com) and that developed by *Trend Micro* (www.trenmicro.com).

There are several programming libraries that permit *pairing* operations over elliptic curves for instance the MIRACL C++ library (http://github.com/CertiVox/MIRACL)and the PBC Library (Lynn 2016) for C, both use the GNU multiple Precision Library GMP. A Java library jPBC (http://gas.dia.unisa.it/projects/jpbc), a port of the PBC library also exists.

3. Identity based cryptography

Identity based cryptography enables information that identifies a user such as an email, name, civil identification number, to be used as a public key thereby eliminating the necessity of key sharing or certificate stores in order to encrypt or sign a message for a given user. In some sense these type of system reduce the overall complexity of public key schemes and also reduce the need for the creation and management of Public Key Infrastructures (PKI).

Figure 1 shows an identity based cryptographic scheme with two communicating entities, Bob and Alice, where Bob encrypts a message to Alice. Alice and Bob both agree to use and to trust a third party called the Private Key Generator (PKG) which is responsible for generating the systems public parameters and for secret key distribution. In general an identity based cryptographic scheme is described by four probabilistic polynomial time algorithms, Setup, Extract/KeyGen, Encrypt, Decrypt

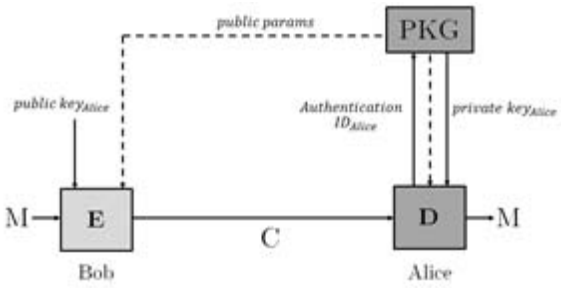


Figure 1: Identity based cryptographic scheme

Setup: In this phase the PKG generates public and private parameters, public parameters can be transmitted to all the users, shown as a dotted line in the figure.

Extract/KeyGen: Here the receiver of the message, in this case Alice authenticates herself at the PKG and obtains *private-key*$_{Alice}$ corresponding to her identity *ID*$_{Alice}$.
Encrypt: Bob encrypts a message *M* using Alice's identity *ID*$_{Alice}$ the encrypted message *C* is sent to Alice.
Decrypt: After receiving the encrypted *C* from Bob Alice decrypts the message using her private key *private-key*$_{Alice}$ and recovers the original message *M*.

Two Identity Based Cryptographic (IBC) schemes were implemented, the *fullident* scheme by Boneh-Franklin (BF), (Boneh 2001) and the Baek-Naini-Susilo (BNS) scheme (Baek 2005) which permits encrypting messages to multiple users. Both of these schemes security properties rely on the hardness of the Bilinear Diffie-Hellman Problem. The next section will briefly describe the BNS scheme, the full details of the BF FullIdent scheme can be found in Boneh (2001).

Notation: Z^+ denotes the positive integers, Z_q is the additive group of integers {0,1,..,q-1} modulo q. For a prime order group G, G^* is used to represent $G\backslash\{O\}$ where *O* is the group identity element.

3.1 The Baek-Naini-Susilo scheme

Setup: Given a security parameter $k \in Z^+$

- Step 1: Generate a prime number q, select two groups G_1 and G_2 of order q, with a bilinear map which satisfies $\hat{e} : G_1 \times G_1 \rightarrow G_2$.
- Step 2: Generate two random numbers $P, Q \in G_1$
- Step 3: Generate random number $s \in Z_q^*$ and calculate $P_{pub} = sP$. The systems master-key is *s*.
- Step 4: Select appropriate cryptographic hash functions H_1 H_2 and H_3

Extract: For the identity string ID calculate $S_{ID} = sH_1(ID)$, Return S_{ID} as the private key of the identity ID.

Encrypt: To encrypt the message M for multiple identities/public keys (ID_1,...,ID_n), generate two random numbers, R and r and calculate the cipher text C such that

$$C = (U, V_1,...,V_n,W_1,W_2,L,\sigma)$$
$$= (rP, rH_1(ID_1) + rQ...,rH_1(ID_n) + rQ,\hat{e}(Q,T)^rR, M \text{ Xor } H_2(R), H_3(R,M,U,V_1, ... ,V_n,W_1,W_2,L))$$

Notice that σ guarantees integrity of the whole ciphertext, L simply contains information about the identities for which it was encrypted. $W_2 = M \text{ Xor } H_2(R)$ can be replaced by $W_2 = E(M)$ where E is a symmetric cypher scheme with key $H_2(R)$. L is simply information that enables recognizing the identities to which the message has been encrypted.

Decrypt: To decrypt the message for a given identity IDi where i = [1..n] :
Given C as (U, V_1, ... ,V_n,W_1,W_2,L, σ), use the information present in L to find the appropriate V_i.
Calculate $R = \hat{e}(U;S_{IDi}) / \hat{e}(Ppub;V_i)\ W_1$, $M = W_2 \text{ XOR } H_2(R)$, and $\sigma' = H_3(R,M,U,V_1, ... ,V_n,W_1,W_2,L))$.
If σ' = σ accept M as the decrypted message else reject

3.2 Efficiency

To finalize this section the efficiency of the two schemes when encrypting a file for multiple end users is compared Table 1 shows the ciphertext size for a message of 32 bytes encrypted for 1 to 50 identities using a 48 byte elliptic curve. Note that using the BF scheme for multiple identities simply implies encrypting the message n times.

Table 1: Ciphertext size

n-identities	BNS Scheme	BF Scheme
1	448 bytes	320 bytes
5	960 bytes	1600 bytes
10	1600 bytes	3200 bytes
25	3520 bytes	8000 bytes
50	6720 bytes	16000 bytes

In terms of runtime efficiency the main cost are the algebraic pairing operations. The following table 2 shows the number of pairing operations when encrypting a file for n-identities. For the BNS scheme the only paring may be pre-calculated and therefore can be considered part of the public parameters.

Table 2: Number of algebraic operations

	Pairings	Add. in G1	Multi. in G2	Exp. in G2
BNS Scheme	0	n	n + 2	1
BF Scheme	n	0	n	n

From these tables we can conclude that even though when encrypting a message for one identity the resulting ciphertext is smaller in the BF scheme the most practical scheme is the BNS scheme since the objective is to encrypt a file and associated security policy that may be accessed my multiple users.

4. Implementation

The final objective is to enable the creation of Desktop applications for the Windows Operating System and also applications for the Azure Cloud platform the .NET C# was chosen as the final target. However since there is no publicity available C# library that enable the use of elliptic curves and the pairing of elements of these curves it was necessary to use a wrapper to a low level C language pairing based cryptographic library, the PBC library, that was used to implement the mathematical operations such as generating elliptic curves and pairing operations. In order to achieve this it was necessary to first create a DLL wrapper that permits the functions in the C library to be called from C# and then to create a library to implement the actual cryptographic schemes

4.1 DLL wrapper

Two wrappers were created for the IBE schemes, BF Fullident and BNS. Each wrapper is divided into two modules, the user or client module and the PKG module, each implements the appropriate elliptic curve operations. Table 3 shows the correspondence between the wrapper functions and the operations necessary for the BNS scheme

Table 3: BNS scheme wrappers for the PKG and client functions

pkg.dll	
Setup	
Generate s (master key)	$s = generateNewMasterKey(curve, (...))$
Generate P parameter	$P = generateNewPparam(curve, (...))$
Generate Q parameter	$Q = generateNewQparam(curve, (...))$
Calculate $P_{pub} = s \cdot P$	$P_{pub} = getPpub(curve, P, s, (...))$
Calculate $\hat{e}(Q,P_{pub})$	$\hat{e}(Q,P_{pub}) = calcpairQPpub(curve, Q, P_{pub}, (...))$
Extract	
Generate $d_{ID} = s \cdot F_{ID_i}$	$d_{ID} = getPrivateKey(curve, F_{ID_i}, s, (...))$

client.dll	
Encrypt	
Generate random element r	$r = genSmallR(curve, (...))$
Generate random element R	$R = genBigR(curve, (...))$
Calculate $U = r \cdot P$	$U = calcU(curve, P, r, (...))$
Calculate $rQ = r \cdot Q$	$rQ = calcrQ(curve, Q, r, (...))$
Calculate $V = r \cdot F_{ID_i} + r \cdot Q$	$V = calcV(curve, r, F_{ID_i}, rQ, (...))$
Calculate $W_1 = \hat{e}(Q,P_{pub})^r \cdot R$	$W_1 = calcW1(curve, \hat{e}(Q,P_{pub}), r, R, (...))$
Decrypt	
Calculate $R = \frac{\hat{e}(U,d_{ID_i})}{\hat{e}(P_{pub},F_{ID_i})} W_1$	$R = decryptCalcR(curve, U, d_{ID}, P_{pub}, F_{ID_i}, W_1, (...))$

4.2 The .Net C# wrapper

The C# libraries developed for key generation, encryption and decryption make use of the previously described wrappers. In order to implement the library to encrypt and decrypt files in a simple and abstract way a Hybrid encryption scheme is used in which a file is encrypted using a 256 bit randomly generated AES key and this key is then encrypted using the BNS scheme described previously (The details are omitted here).The library enable any number of final applications may be developed and consists in functions for systems setup, key generation and encryption/decryption.

The PKGs function is to generate the system parameters, the systems private and public parameters and also the private key for a given identity. This is achieved using the following library functions :

byte[] GenerateMasterKey(string curve)
byte[] GeneratePparam(string curve)
byte[] GenerateQparam(string curve)
byte[] CalculatePpubparam(string curve, byte[] Pbyte, byte[] masterKey)
byte[] CalculatePairQPpub(string curve, byte[] Qbyte, byte[] Ppubbyte)

These functions receive as parameter a string which contains a description of the chosen elliptic curve and hence only need to be executed once. After generation the systems master or private key s should remain private as the security of the overall system depends on this. The public parameters and elliptic curve can be made publicly available to the users of the system.

To generate a private key (extract) the necessary parameters are the users identity, a string, the elliptic curve and the systems master key.

byte[] GetPrivateKey(string id, string curve, byte[] masterKey)

The user can call C# functions to encrypt and decrypt file streams in a relatively simple fashion without knowing the internal workings of the underlying schemes using the functions *EncryptStream* and *DecryptStream* as shown below.

MemoryStream EncryptStream(string[] identities, string policy, MemoryStream streamIn, string originalFileName, string curve, byte[] paramP, byte[] paramQ, byte[] parameQPpub)

MemoryStream DecryptStream(string identity, byte[] privateKey, MemoryStream streamIn, string curve, byte[] paramPpub)

The encryption function receives as parameters, the identities of the receivers, an optional string containing additional policy information, the stream to encrypt, an optional parameters containing the original file name to be encrypted, the elliptic curve and public parameters P, Q and $\hat{e}(Q,P_{pub})$.

The decryption function receives the users identity and corresponding private key, the stream to be decrypted, the elliptic curve and public parameter P_{pub}.

In case of error these function throw exceptions with relevant error information.

The relevant library code is available online at https://spocs.it.ubi.pt/price/ibc.html

4.3 Sticky policies

The concept of sticky policies was introduced by Karjoth et al. (2002). A sticky policy is a security policy that is attached to data or a data file and which describes the security and privacy policies that a user accessing this data must adhere to. This type of mechanism enforces a fine-grained access and control mechanism on the data. The Sticky policies architecture used in this work is similar to that described by Mont et al. (2003). In this case users who can access the encrypted files must also respect the files security policy before the PKG supplies a private key in order to decrypt the file, hence in this case the PKG is also the policy enforcer

Various types of policy may be imagined, temporal policies, role based, IP and MAC address etc. In our scheme two policies were defined, a temporal policy called DATE and a role based policy security level called SECURITY

The DATE element has two parameters, Start and End, and defines a time period during which a user may obtain a private key from the PKG. For the SECURITY policy the PKG must have or have access to a database in order to match a given users ID with a particular security level. For this proof of concept four security levels were defined. Hence for a given files security policy the PKG must first look up the users identity and check that the users has a sufficient security level in order to access the file

```
<?xml version="1.0" encoding="utf-8"?>
<POLICY>
  <DATE active="true">
    <Start>11/01/2013-00:00</Start>
    <End>01/01/2014-00:00</End>
  </DATE>
  <SECURITY>
    <Type>Level 1</Type>
  </SECURITY>
</POLICY>
```

Figure 2: Example of a XML policy file

Figure 2 illustrates an example of one of the XML policy files is given. In this case when a user requests a private key to access the file, the users identity must be one of the identities to which the file has been encrypted using the BNS scheme and also the PKG can check that the current data is between the first of November 2013 e the first of January 2014 and also that the user has level 1 security or higher.

Notice that the security policy is first defined by the data owner and then attached to the file when it is encrypted. The access is enforced when a user requests a private key from the PKG o access the files contents.

In order to associate a policy file to a data file a hash is taken of the policy file and this is then combined with the users identity in order to create a public encryption key **identity + hash**.

The combination of Hash and Identity creates a connection between the polices and the encrypted file, since the hash used can only be generated from the original policies. If an attacker tries to alter the policies the private key generated will bit decrypt the file, note that in this way the policies must be sent to the PKG together. Hence it's convenient to define a format for the encrypted file which also includes the policy file. We have defined an appropriate format for files that have been encrypted to multiple identities only whose default extension is .ibc and for files that have been encrypted to multiple identities with an associated policy we have a used a zipped file container with extension .ibz which contains the xml policy file and the encrypted file, this is shown in Figure 3.

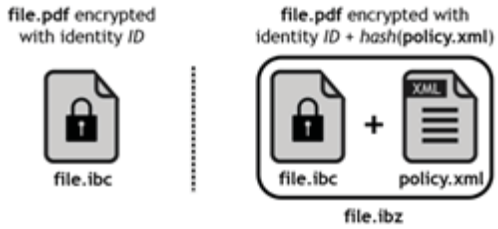


Figure 3: File encryption

5. Applications

This section will describe the applications that were developed for file encryption using the C# library and the file formats specified previously. The complete description can be found in (Silveira 2013)

5.1 Desktop application

A Windows client desktop application in C# was developed and is shown in (Figure 4), the PKG runs as Cloud service using the Windows Azure platform.

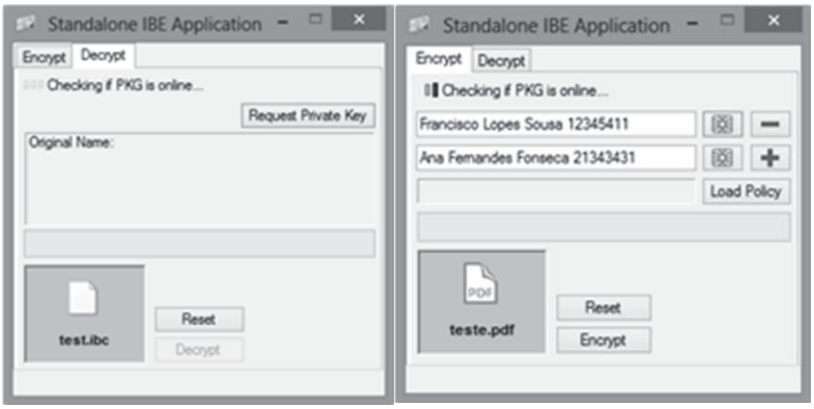


Figure 4: Windows client application

The application permits the encryption/decryption of files to multiple identities. The identities are derived from characteristics present in a E- ID card, in this case the Portuguese Citizens Card. The characteristics used as a users public key are the *Full Name* and *Civil Id number*. In order to obtain a private key its necessary to authenticate at the PKG using the E-Id cards' authentication certificate and authentication mechanism (pin or biometric).

To ensure system security, all communications between the application and the PKG service are made over Secure Sockets Layer (SSL/TSL) sessions. A session begins when the client executes the application and this connects to the PKG and ends when the user receives a private key, error message or after session inactivity timeout.

To decrypt a file the client application has to request a private key from the PKG using the E-Id's authentication certificate, Figure 5 shows this process which is described below

- 1. The application sends the CC authentication certificate to the service, the identity of the user, the policy file and the policy files *hash*.
- 2. The PKG cloud service after receiving the data makes the following verifications
- *i.Confirm that the certificate send contains the identity of the requesting user;*
- *ii.Confirm that the certificate is indeed a valid authentication certificate of the relevant PKI, by checking the certificate chain and checking if the certificate has been revoked;*
- *iii.Verify that the hash of the policy file is the same as the hash sent;*
- *iv.Verify the XML Policy File and apply/verify the polices.*

If any of these conditions fails then the (SSL) session is terminated and the application receives an error message.

- 3. The service generates and returns a 128 bit *Globally Unique Identifier* (GUID)
- 4. The client side user signs the GUID with the private key present in the private authentication certificate on the E-ID Card(the CC requests the client authentication pin) and then sends this signature to the PKG service.
- 5. PKG verifies if the signature is valid using the public key in the clients public authentication certificate;
- 6. If the signature is valid the PKG generates and sends the private key.

The service does not keep any permanent information about users or store users keys. Both the certificate and the user's identity are discarded as soon as the session is over. The fact that computing the encryption and decryption of files is made on the client side allows the load carried on the cloud the service to be fairly light.

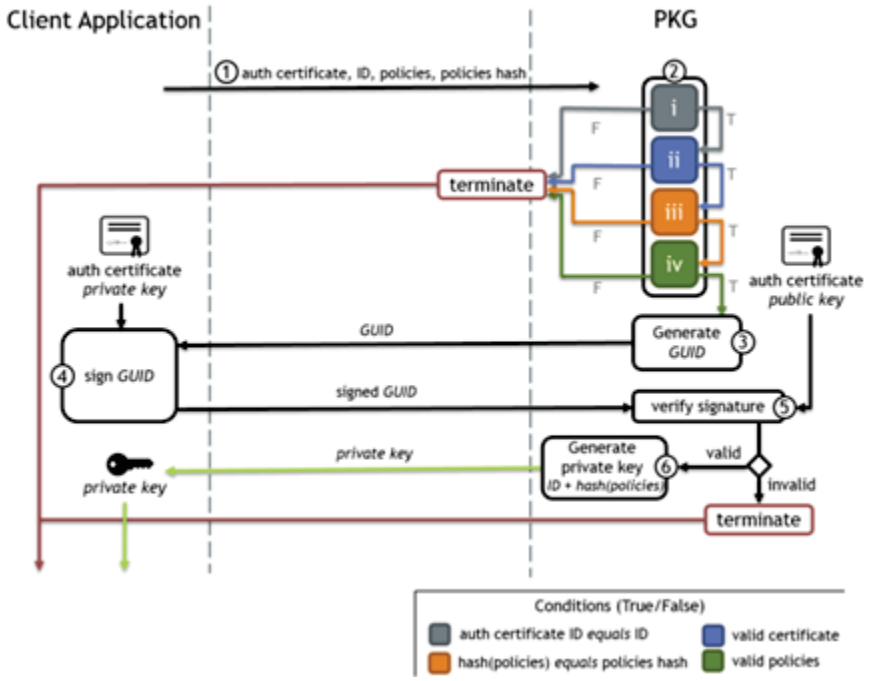


Figure 5: Private key request process

5.2 Web application

In addition to the Desktop application the PKG service is also being used in a *web* application for *Cloud* storage which permits encryption and decryption of files also using the Citizen Card (Figure 6). The PKG allows concurrent access so that multiple Desktop or Web applications can access the service simultaneously.

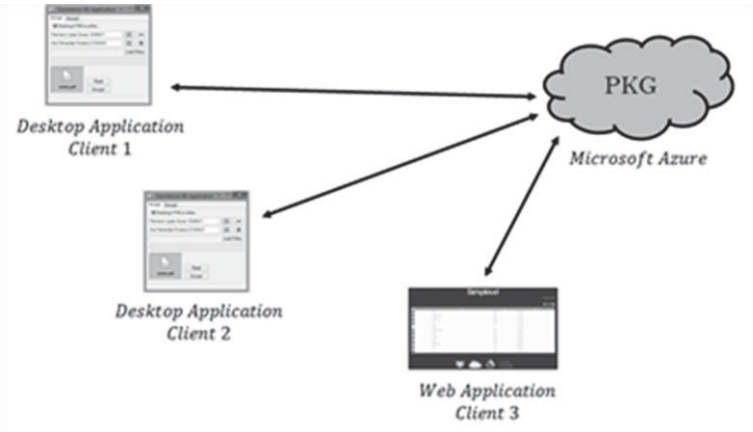


Figure 6: Web and cloud applications

6. Conclusions and future work

We have briefly described an innovative encryption scheme that enables the use E-ID cards that contain strong authentication and digital signature mechanisms but no standard encryption keys using the publically availably identity features on these types of cards. We have also developed C# wrappers for the PBC pairing library and shown that the proposed systems a provides a secure system for file confidentiality and privacy as well as being transparent and easy to use by the end users of the system. Future work that would be relevant to the present system would be to use threshold encryption for dividing the systems master key in several PKG services, hence decentralizing the power of one PKG .

Acknowledgements

The authors would like to thank the support of the Instituto deTelecomunicações, the Fundação para a Ciência e Tecnologia (FCT) Portugal UID/EEA/50008/2013, the Soft SIM Project Portugal Telecom and the RELEASE laboratory of the University of Beira Interior.

References

Baek,J, Safavi-Naini, R, and Susilo (2005) W. *Efficient multi-receiver identity-based encryption and its application to broadcast encryption*. In Serge Vaudenay, editor, Public Key Cryptography - PKC 2005, volume 3386 of Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Boneh, D and Franklin, M. (2001) *Identity-based encryption from the weil pai*ring. In Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology, London, UK. Springer-Verlag.

Boneh, D, Gentry, G, and Hamburg, M. (2007). *Space-efficient identity based encryption without pairings*. 48th Annual IEEE Symposium on *Foundations of Computer Science*.

Cocks, C. (2001) *An identity based encryption scheme based on quadratic residues.* In Bahram Honary, editor, Cryptography and Coding, volume 2260 of Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Karjoth,G, Schunter M, and Waidner (2002) M. *Platform for enterprise privacy practices: Privacy-enabled management of customer data*. pages 69,84. Springer.

Lynn, B. *Pbc library, the pairing-based cryptography library*. http://crypto.stanford.edu/pbc/.

Mont, M, Pearson S, and Bramhall P. (2003) *Towards accountable management of identity and privacy: Sticky policies and enforceable tracing services*. pages 377,382. IEEE Computer Society.

Silveira, J (2013) Aplicações de Criptografia Baseada em Identidade com Cartões de Identificação Electrónica. MSc Thesis, University of Beira Interior Portugal.

Shamir A (1984) *Identity-based cryptosystems and signature schemes*. Advances in Cryptology: Proceedings of CRYPTO 84 Lecture Notes in Computer Science Volume 196, 1985, Springer Berlin Heidelberg.

Security Implications of SCADA ICS Virtualization: Survey and Future Trends

Tiago Cruz, Rui Queiroz, Paulo Simões and Edmundo Monteiro
University of Coimbra, Portugal
tjcruz@dei.uc.pt
rqueiroz@dei.uc.pt
psimoes@dei.uc.pt
edmundo@dei.uc.pt

Abstract: In recent years, Supervisory Control and Data Acquisition (SCADA) Industrial Control Systems (ICS) – a kind of systems used for controlling industrial processes, power plants or assembly lines – have become a serious concern because of security and manageability issues. Years of air-gaped isolation, the increased coupling of ICS and Information and Communication Technology (ICT) systems, together with the absence of proper management and security policies, disclosed several weaknesses in SCADA ICS. Suddenly, these systems were faced with a reality that was familiar for ICT infrastructure managers for decades, which has driven the need for the development of specific technologies, as well as the establishment of management frameworks and the adoption of security-oriented policies. Virtualization was one of such developments, whose influence spawns several domains, from networking and communications to mass storage and computing resources. For ICT, the rise of virtualization constituted a paradigm shift, with significant gains in terms of resource consolidation, manageability or even security. These benefits are yet to fully reach the ICS domain, despite recent developments geared towards the introduction of hypervisors or software-defined networking within such systems. This paper provides an overview on the usage of such technologies to improve SCADA ICS security and reliability also proposing advanced use cases.

Keywords: virtualization, critical infrastructure protection, industrial control systems

1. Introduction

In recent years, SCADA ICS – a kind of systems used for controlling power plants, assembly lines or industrial processes, often part of critical and/or strategic infrastructures – have become a serious concern because of security and manageability issues. After years of air-gaped isolation, the increased coupling of ICS and ICT systems, together with the absence of proper management and security policies (Krutz 2006), disclosed several weaknesses in SCADA ICS, which were left exposed to attacks, with potentially catastrophic consequences. Nevertheless, these problems hardly constitute any novelty within the ICT domain, which has dealt with them for decades, driving the need for the development of specific tools and protocols, as well as the establishment of management frameworks, such as Information Technology Infrastructure Library (ITIL) change management (Gallup 2009) or security oriented policies.

However, ICT-specific practices cannot be easily ported to the ICS domain. For ICS operators, equipment manufacturers and software developers alike, reliability is top priority. Continuous operation and operational safety targets make it difficult to deploy several ICT-specific strategies and tools, because of the potential impact on the ICS. This has pushed the industry, researchers and standardization organizations to conceive ICS-specific security and management solutions and frameworks, as well as publishing guidelines and guides documenting best practices. New product lines were also introduced, with added security features and management capabilities.

Still, the ICS paradigm itself remained relatively unchanged, as proposed solutions try to fix what is wrong without attempting to introduce significant change into existing systems. This solution is far from optimal, as typical lifecycle management operations such as security patch deployment are still an issue in modern SCADA ICS, the same being true for change management. In contrast, these issues have been addressed in the ICT domain for years, through the continuous development of technologies, tools and practices, designed to address such needs. Virtualization technologies are among these developments, which influence ICT computing and communications infrastructures. Developments such as hypervisors, Software-Defined Networking (SDN) or Network Function Virtualization (NFV) are reshaping the ICT ecosystem, providing the means to rationalize the use of computing and communications resources, also being instrumental to optimize and/or improve aspects such as lifecycle management, energy efficiency, reliability or security, among others.

From an ICS security and reliability perspective, device and infrastructure virtualization may have a similar impact as they had for ICT, as the industry slowly starts to absorb some of the technologies, customized and fine-tuned for critical infrastructure environments. However, this is a process undergoing its early stages, not only because the specific ICS use cases for several virtualization technologies have yet to be developed, but also because extensive testing is required for its certification in such environments. In this scope, this paper analyses the application of virtualization technologies for communications and computing resources in ICS contexts, with a focus on recent developments, open challenges and benefits, from a security and reliability-oriented perspective.

The rest of this paper is structured as follows. Section 2 discusses the problem of security in ICS/SCADA, also explaining the potential benefits of introducing domain-aware virtualization technologies in such environments. Section 3 discusses the introduction of network virtualization technologies in SCADA ICS and its security benefits. Section 4 addresses the advantages of introducing partitioning hypervisors in ICS, describing a virtualized Programmable Logic Controller (PLC) use case. Finally, section 5 presents conclusions insights about future developments.

2. Virtualization and SCADA ICS security

As their scope was originally restricted to isolated environments, SCADA systems were considered relatively safe from external intrusion. However, as architectures evolved, these systems started to assimilate technologies from the ICT world, such as TCP/IP and Ethernet networking. This trend, together with the increasing adoption of open, documented protocols, exposed serious weaknesses in SCADA architectures, a situation that was aggravated by factors like the use of insecure protocols, such as Modbus (Triangle 2002) or inadequate product lifecycle management procedures (Igure 2006), the latter being responsible for the proliferation of devices and components beyond their end-of-life support status. Also, the interconnection of the ICS network with organizational ICT network infrastructures, and even with the exterior (for example, for remote management) brought a new wave of security incidents, with externally initiated attacks on ICS systems increasing significantly, especially when compared with internal attacks (Kang 2011). Overall, this situation has become the root cause of many well-known ICS security incidents, such as the Stuxnet Trojan (O'Murchu 2011).

In fact, ICS security cannot be approached in the same way as its ICT counterpart, as both domains differ significantly on their fundamental design principles. Due to its critical nature, ICS operation and design practices frequently privilege availability and reliability over confidentiality and data integrity – a perspective that is quite the opposite from the ICT philosophy, which follows an inverse order of priorities (ISA-99.00.01).

The differences between the ICT and ICS domains also mean that there is no "one size fits all" solution when it comes to choose and implement security mechanisms. The fundamental premises for ICT security tools and commonplace lifecycle management procedures, such as patching and updating a system, can become troublesome in an ICS, when faced with situations such as the impediment / high cost of stopping production (Zhu 2011), or even the explicit prohibition by the system's manufacturer, as any software release has to be certified before being released. Also, several security mechanisms, such as anti-virus software are frequently unadvised by SCADA software providers, as they might interfere with the response latency of the host. The same rationale applies to anything deployed in the middle of the critical communications path (e.g., an inline network Intrusion Detection System), as it may induce latency or some other sort of reliability issue.

Ironically, much of the problems faced by ICS are not entirely new, as they were known well before in the ICT domain, which has undergone several paradigm shifts and major technological steps to deal with them. More recently, the rise the virtualization paradigm has become instrumental in changing the ICT computing landscape, providing the means to leverage computing and communications resources, through consolidation and efficient management. Technologies such as hypervisors, SDN or NFV are contributing to rationalize, streamline and reshape infrastructures and devices, up to the point of changing the way communications and computing resources are consumed by end-users.

In terms of security and reliability, the impact is manifold. For instance, by creating a virtual machine (VM) snapshot it is possible to rollback changes in case of failure or corruption caused by a failed OS patch or malicious tampering; VMs can be cloned for sandboxed testing, prior to deployment into production; hypervisors can perform in-place behavior monitoring of instances for security and safety purposes. Similarly, technologies such

as SDN, which constitute a flow-oriented virtualization mechanism for networks, allow for the flexible creation and management of network overlays on top of existing physical infrastructures, while also enabling significant security and reliability benefits (Proença 2015). NFV, in its turn, can work together with SDN to virtualize network equipment functionality, spreading it across the communications and computing infrastructure in an efficient and rational way, also enabling the creation of innovative security solutions designed to better couple with the increasingly distributed nature of modern ICS and associated threats (Cruz 2015).

But the introduction of ICT-like virtualization techniques in ICS is not a straightforward process. For operators, equipment manufacturers and software developers alike, reliability, operational safety and continuous operation are top priorities, a situation that makes it difficult to deploy several IT-specific strategies and tools, because of the potential impact on the ICS. For example, the latency overhead of certain mechanisms may not be compatible with real-time operation requirements. Hypervisors must cope with the (soft) real-time requirements of ICS applications; any attempt to introduce SDN or NFV must account for the potential impact in terms of ICS reliability or latency.

Despite the constraints, the potential efficiency, security and reliability benefits for ICS are enough to justify the progressive development and introduction of domain-aware virtualization technologies. For instance, real-time hypervisors can provide safe partitioning and isolation, enabling the creation of managed execution environments for real-time workloads, with continuous assessment of partition behavior, also providing rollback capabilities for potentially compromised systems. Use of SDN technologies can provide the ICS operator with the means to monitor the ICS communications infrastructure behavior, while easing the implementation of countermeasures and deployment of security mechanisms. As ICS become increasingly distributed, NFV can provide the means to efficiently spread functional security components across the ICS communications and computing infrastructure, in order to better couple with the dispersed nature of the protected systems. The next two chapters will discuss how domain-aware virtualization can provide effective security benefits for ICS, with a focus on two major scopes: communications and computing.

3. Virtualization of SCADA ICS communications infrastructures

This chapter is specifically concerned with the introduction of SDN and NFV technologies within the SCADA ICS scope. For this purpose, the security benefits of the technologies hereby discussed will be analyzed from a broad perspective, both in terms of the physical ICS dimension and dispersion of its scope, ranging from plant-level to distributed Industrial Automation and Control Systems (IACS) use cases. All sections will start with a brief introduction of its respective cornerstone concepts, namely SDN and NFV, in order to ease its introduction in the context of SCADA ICS security.

3.1 SDN and SCADA ICS

SDN is an architecture that decouples forwarding functions (data plane) and network control (control plane), with the aim of introducing direct programmability into the network, to applications and policy engines alike. With SDN, packet forwarding is flow oriented, meaning both origin and destination are taken into account, instead of just packet destination, as in traditional networking. Flow policies are granted by an SDN controller, which manages the policies for a range of forwarding elements in a given network, effectively moving control plane functions outside of the devices. Thus, SDN-capable elements can be dynamically reconfigured over the network accordingly with the needs of network services and applications. For this reason, the controller will have a broader view of the domain, contrasting with the narrow view that an individual forwarding element has in a traditional IP network. There are several SDN protocols, among which OpenFlow (ONF) is one of the most popular. SDN allows for increased network flexibility and programmability, in particular for complex scenarios, which benefit from the reduced overhead for management operations such as topology changes for implementing overlay networks. Besides these benefits, SDN can also provide an effective mechanism for security applications (Proença 2015). This is due to the fact that a centralized element with a global view of all the network entities – such as devices, flows and network elements – is able to provide more efficient information gathering and security reaction mechanisms, especially when compared with the narrow local view individually provided by each forwarding element in traditional IP networks. Moreover, flow-based forwarding can be used to increase the efficiency of a reaction, being used to isolate or divert flows, instead of simply blocking an attack. This is useful to improve existing security techniques – for example, allowing to dynamically divert attackers to honeypot systems, as soon they are detected. SDN can also help handling Denial of Service (DoS) and Distributed DoS (DDoS) attacks, by improving detection and reaction mechanisms.

Besides the generic security application scenarios, there have been several developments regarding SDN-based security mechanisms for ICS. For instance, (Dong 2015) proposes reinforcing the resilience of SCADA networks used for smart grid applications using a solution relying on three elements (SCADA master, SDN controller, Intrusion Detection System – IDS), which coordinate with each other in order to detect attacks and reconfigure the network so as to mitigate and overcome identified problems. Suggested use cases include the dynamic establishment of routes to transmit control commands only when necessary (to shorten the time window for tampering attempts); automatic rerouting or dropping of suspicious packets to avoid spoofing or flooding attacks from compromised SCADA elements; or the implementation of network monitors to deal with delay attacks.

(Irfan 2015), proposes using SDN for dynamic creation of virtual networks in order to isolate distinct traffic and hosts, enabling traffic prioritization and secure partitioning. The concept is demonstrated using an SDN controller proxy to create three isolated networks, which share the same physical infrastructure, but have their own SDN controllers. Authors discuss the use of this architecture to improve aspects such as authentication, confidentiality, integrity, non-repudiation and availability. A similar approach is also suggested by (Machii 2015) as a way to minimize the attack surface, by using SDN to dynamically segregate fixed functional groups within the ICS. A dynamic zone-based approach is also proposed, taking advantage of the information obtained from field devices to estimate the operation phase of the ICS (as each phase, such as start-up, normal operation or load-change exhibit different behavior and communications profiles) and calculate the optimal zone topology, deploying the needed SDN configuration in runtime. This strategy reduces the time and spatial exposure to attacks, also providing the means to isolate compromised devices.

Also related to dynamic configuration techniques, (Chavez 2015) presents a security solution based on network randomization, complemented with an IDS capable with near real time reaction capabilities. This network randomization approach assigns new addresses to network devices in a periodic basis or by request, in order to protect them against attacks that rely on knowledge about the ICS topology (such as static device addresses). The responsible controller application keeps an updated database of all the network specifications (mostly devices and real addresses), generating overlay IP addresses for the same devices and for each flow, which are used to define the OpenFlow rules on flow tables. This way, all the traffic flowing on the network uses “fake” overlay addresses that are periodically randomized, reducing their useful lifetime and, consequently, the time window available for any attacker to take advantage of that knowledge. The proposed IDS takes advantage of the predictable, auto-similar, traffic patterns of ICS networks for identify attacks and trigger defense reactions (a network randomization request, which will render useless any ongoing attack using old overlay addresses). Attack detection makes use of machine learning algorithms and mathematical methods, fed and trained using OpenFlow’s statistical counters. (Silva 2015) also describes a dynamic technique that makes use of SDN to prevent eavesdropping on SCADA networks. The intended goal is to deter attackers from collecting sequential data, which is essential for breaking encryption, identify patterns and retrieve useful information from the payload. By taking advantage of redundant network connectivity, a multi-path routing mechanism enables a flow to be transmitted and split over different paths (see Figure 1) by resorting to an algorithm that calculates the shortest path between two devices, dynamically assigns a cost to each one and uses an OpenFlow timer (hard timeout) to periodically reinstall new flow rules.

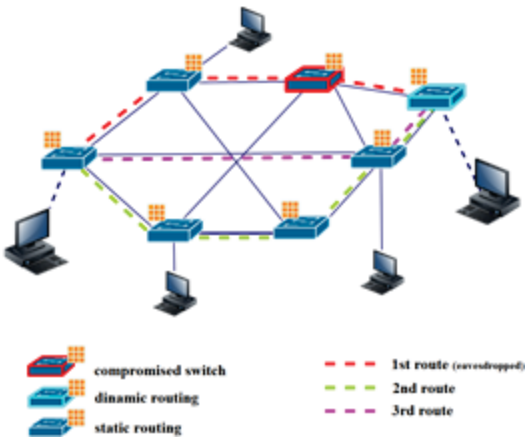


Figure 1: Multi-flow, redundant routing for flow splitting (adapted from (Silva 2015))

(Genge 2016) proposes two distinct SDN-based techniques to mitigate and block ICS cyber attacks. The first technique (see Figure 2), designed for single-domain networks, attempts to mitigate DoS attacks by rerouting traffic, using information from the SDN controller. SDN controllers feed an application that continuously monitors the state of the network links and communicates with the controller to issue flow reconfiguration operations. Once an attack is detected (few details are provided about this, though), the corresponding data flows are rerouted, in order to protect the ICS.

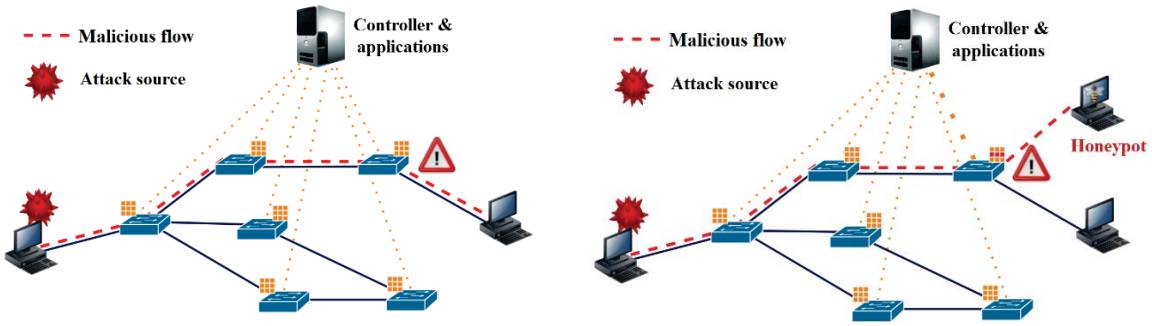


Figure 2: A single-domain SDN-based security solution (adapted from (Genge 2016))

The second technique (see Figure 3) targets multi-domain networks, with the goal of blocking the attack as close as possible to the entry point in the network.

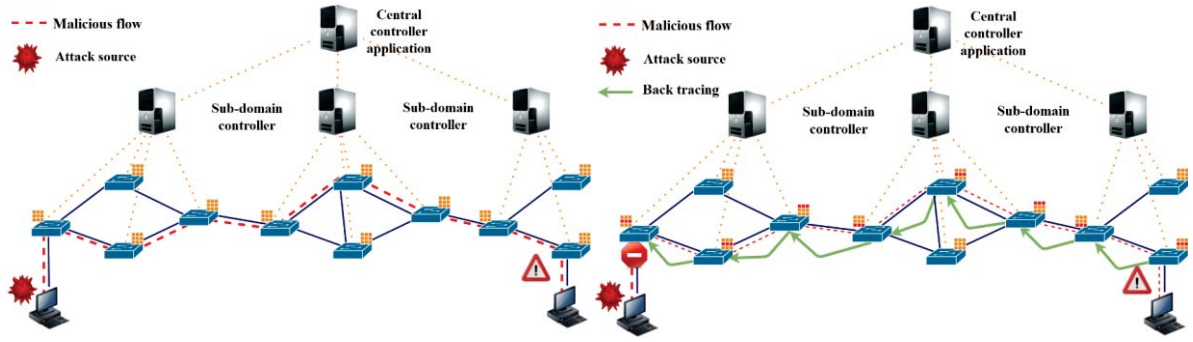


Figure 3: A multiple-domain SDN-based security solution (adapted from (Genge 2016))

For such a multi-domain network, each domain has its own OpenFlow controller, connected to a centralized security application. This application receives information from the SDN controllers, having access to a global perspective about the network – once an attack is detected, it will backtrack towards its origin, by recursively issuing queries about the related flows to identify the previously paired nodes, until the original network entrance point is found.

3.2 Network function virtualization and distributed ICS

NFV is the result of the convergence between telecommunications infrastructures and infrastructure virtualization. As network applications and services scale and evolve (not only in sheer capacity requirements, but also in complexity), they imposed an added burden to the supporting telecommunications provider infrastructure, requiring the use of specific network management and traffic policies that cannot be provided by the network. In this perspective, NFV (Chiosi 2012) is a significant development as it enables the creation of flexible and on-demand network services through a service chain-based composition mechanism that uses network functions implemented in VNF (Virtualized Network Functions) components comprising functionality such as NAT, IDS, Firewalls or other service modules, implemented as VM appliances. The NFV vision attempts to decouple network capacity from functionality, by conceiving an end-to-end service as an entity that can be modeled and described by means of network function forwarding graphs (Figure 4) involving interconnected VNFs and endpoints (also known as service chaining.

Figure 4: NFV forwarding graph example

This approach allows for creation of differentiated end-to-end services that can be provided by the (ordered) combination of elementary VNF or physical functions, chained together by a Forwarding Graph, which models the service flows (see Figure 5). Furthermore, VNF FGs can be nested to define complex functions. VNFs are implemented in software, being interconnected through the logical links that are part of a virtualized network overlay, which can be implemented using SDN.

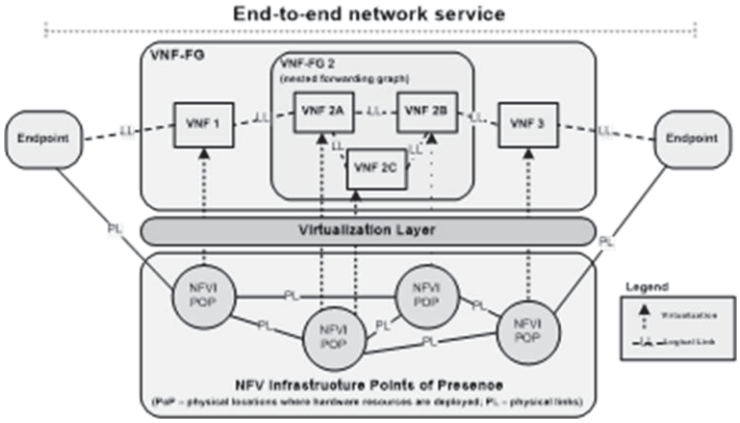


Figure 5: NFV end-to-end service with VNFs (adapted from (Ersue 2013))

Eventually, even Physical Network Functions (conventional network devices with close coupled software and hardware that perform network functions) can be involved in a Network Forwarding Graph service chain (the concept of service chain is not exclusive of NFV). A virtualization layer abstracts the physical resources (computing, storage, and networking) on top of which the VNFs are deployed and implemented, with the supporting NFV Infrastructure (NFVI) being spread across different physical locations, called Points of Presence (NFVI PoPs), as shown in Figure 5.

3.2.1 NFV as an enabler for a new generation of distributed IACS

Use cases such as Internet of Things (IoT), wire to water generation, micro generation, smart metering or smart water management constitute a new generation of distributed IACS that can only be supported with the help of a complex distributed software stack, potentially also requiring the involvement of third-parties, such as telecommunications and cloud operator infrastructures – for this reason, the introduction of Network Function Virtualization component appliances, distributed across geographically dispersed infrastructure PoPs, makes entire sense,.

As the IACS enters the customer premises, the NFV service abstraction model (services as composition of VNFs) provides an effective way to introduce support components along the service path – for instance, a data collection and analysis VNF can be added to the customer service chain (eventually within a virtual Business Gateway service abstraction) to provide data collection for smart metering scenarios. The same rationale applies for security purposes, as cyber-physical protection (for example, to implement bump-in-the-wire encryption) or security anomaly detection VNFs can be integrated within service chains, also using SDN to create flexible security monitoring and reaction capabilities. Moreover, Distributed IDS (DIDS) components may be consolidated in the form of VNFs optimally deployed in order to reduce service overhead and rationalize resources. For instance, the DIDS components might be deployed in the form of VNFs, either shared among several Business Gateway FGs or used exclusively by a service instance (Cruz 2015). Some manufacturers (RAD 2015) (ECI 2015) are starting to propose NFV products for ICS applications that implement this philosophy, NFV capabilities in access nodes for optical transport or packet switched networks, for hosting firewall, encryption or traffic monitoring VNFs

NFV is also an enabler for fog computing (or “edge computing”) scenarios, allowing parts of the infrastructure to be deployed on the network edges, using virtualized platforms located between end user devices and the cloud data centers. This approach addresses the need to process large data streams in real time while working within the limits of available bandwidth, by placing some of transactions and resources at the edge of the cloud (in locations close to end users), thus improving the efficiency of the infrastructure by offloading processing tasks before passing it to the cloud. For these reasons, fog computing is becoming a cornerstone concept for

distributed IACS architectures, providing a way to deal with the information volume generated by sensor streams in an efficient way.

The NFV paradigm is naturally compatible with fundamental premises for implementation of fog computing distributed topologies. As such, it is envisioned that distributed awareness and IACS cyber-security detection capabilities will take advantage of the NFV paradigm to support its underlying deployment model, departing from the conventional, self-contained model and moving towards an architecture capable of keeping up with the geographically dispersed nature of IoT IACS. Also, the VNF deployment criteria may consider the availability of specific capabilities (such as raw processing capacity) in a specific NFVI POP – for instance, per-subscriber security event processing components may be hosted in a different NFVI POP from the one(s) hosting other VNFs for the DIDS service.

4. Real-time hypervisors + SDN = towards a virtualized PLC

Born in the mainframe era, Virtual Machine Monitors (also called Hypervisors) have ultimately evolved towards being supported in open, Commercial Off-the-shelf (COTS) hardware platforms. Specifically, type-1 (bare metal) hypervisors have become popular in large-scale virtualization scenarios such as datacenters, bringing several benefits in terms of resource consolidation, business continuity, scalability, management and security.

But most type-1 hypervisors are optimized for ICT loads, being unsuitable for several ICS application use cases, mostly due to the overhead of the mediation and translation mechanisms abstracting the host hardware from the VM. This situation gradually began to change, as some operators started virtualizing hosts with services deployed on general-purpose OS, such as SCADA Master Stations (MS), Human-Machine Interfaces (HMI) or Historian Database servers (HDB), using conventional type-1 hypervisors. This was possible due to the development of hardware-assisted memory management and I/O mechanisms to implement robust resource affinity and reservation (such as VT-d and PCI SRV-IO (Garcia-Valls 2014) support), providing performance guarantees while avoiding the effect of resource overprovisioning.

Other ICS elements, such as process control devices can also potentially benefit from virtualization technologies. For instance, (Cahn 2013) proposed the virtualization of Intelligent Electronic Devices (IEDs) used to collect information from sensors and power equipment, with the purpose of optimizing the maintenance and cost overheads, while increasing reliability. The same rationale could be applied to Programmable Logic Controllers (PLC) devices, which constitute the focus of this section.

PLCs are pervasive devices in ICS, such as SCADA systems, being designed to control industrial processes. Contemporary PLCs are the outcome of an evolutionary process that started with the first generation of relay-based devices, progressively incorporating technologies such as microprocessors, microcontrollers and communications capabilities, ranging from serial point-to-point or bus topologies to Ethernet and TCP/IP. Despite modern PLCs often being embedded devices with commodity Instruction Set Architecture (ISA) System-on-Chip or CPUs (PowerPC, x86 or ARM), running Real-Time Operating Systems (RTOS), its virtualization was not deemed feasible until recently, due to the lack of specific hardware, software and infrastructure support.

4.1 Towards the virtual PLC

PLCs are designed for reduced and deterministic latency, operating under strict timing constrains that are dependent on factors such as the end-to-end and event response latencies across components on interconnected buses, or signal and message propagation delays. These requirements are incompatible with the use of several virtualization technologies, such as conventional type-1 hypervisors, due to overhead issues and the lack of support for real-time payloads.

However, recent developments such as the implementation of low-latency deterministic network connectivity for converged Ethernet and the availability of real-time hypervisors made it possible to virtualize components of the PLC architecture. The vPLC architecture hereby proposed (Figure 6) takes advantage of these capabilities, by decoupling the PLC execution environment from I/O modules – using an SDN-enabled Ethernet fabric to provide connectivity to the I/O subsystem. This architecture departs from the SoftPLC concept, as proposed by products such as (Codesys) or (ISaGRAF), by adopting an approach in line with (Intel 2013) and (IntervalZero 2011), with the added benefit of a convergent fabric scenario with SDN capabilities.

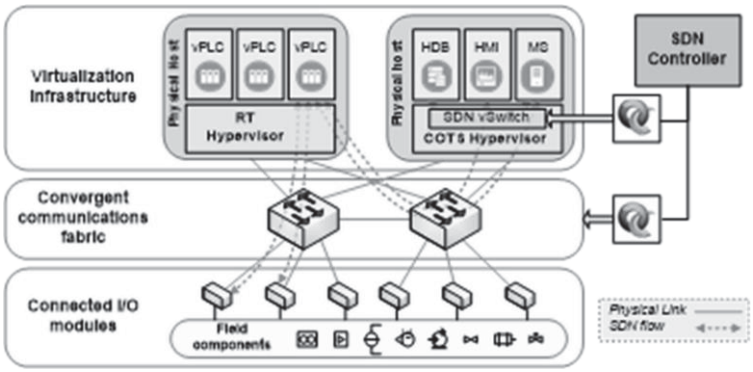


Figure 6: The vPLC architecture

In the vPLC, the PLC I/O bus is replaced by high-speed networking capabilities, with SDN allowing for the creation of flexible virtual channels on the I/O fabric, accommodating the connectivity flows between the vPLC instances and the I/O modules, such as sensor interfaces or motion controllers, providing traffic isolation. Moreover, such I/O modules can be built with reduced complexity, thanks to recent progress in terms of Field-Programmable Gate Arrays (FPGA) and Application Specific Integrated Circuit (ASIC) technology. SDN reconfiguration is managed by means of an SDN controller, via a High-Availability (HA) server (not depicted in the figure), which interacts with its northbound interface. The HA server continuously monitors the SDN switch statistics and path reachability, triggering reconfiguration procedures in case of performance degradation or failure.

This decentralized model shares similarities with remote or distributed I/O PLC topologies, with networked I/O modules acting as extensions of the PLC rack. This goes in line with the Converged Plantwide Ethernet (CWpE) (Didier 2011) architecture, or even critical avionics systems, which replace legacy interconnects with Ethernet-based technologies, such as Avionics Full-Duplex Switched Ethernet (AFDX) (Fuchs 2012).

Advances in cut-through switching, together with Remote Direct Memory Access techniques (RDMA), particularly in converged Ethernet scenarios, have allowed for port-to-port latencies of the order of the hundredths of nanoseconds in 10G Ethernet switch fabrics and application latencies in the order of microseconds (Beck 2011). Additionally, resources such as Intel's Data Plane Development Kit (DPDK) (Zhang 2014) allow for the implementation of low latency, high-throughput packet processing mechanisms that bypass kernels, bringing the network stack into user space and enabling adapters to perform Direct Memory Access operations to application memory. This enables satisfying requirements for single-digit microsecond jitter and restricted determinism, allowing for bare-metal performance on commodity server hardware. On top of this, proposals such as the 802.1Qbv Time Sensitive Networking (IEEE) standard provide compliance with real-time requirements in the microsecond range on conventional Ethernet.

As for computing resources, there are two factors that must be considered. First, modern x86 or ARM processors have become capable of replacing microcontrollers in standalone PLC applications (Kean 2010), due to improvements in terms of raw performance, low latency I/O mechanisms or the availability of ISA extensions suitable for Digital Signal Processing tasks. Second, the availability of real-time static partitioning hypervisors, such as Jailhouse (Siemens), Xtratum (Crespo 2010), X-Hyp (X-HYP) or PikeOS (Baumann 2011) enables hosting RTOS guest VMs for real-time workloads. Some hypervisors, such as Xtratum and PikeOS, even replicate the ARINC 653 (Fuchs 2012) partitioning model for safety-critical avionics RTOS, with a Multiple Independent Levels of Security/Safety (MILS) (Alves-Foss 2006) architecture.

The benefits of this approach are manifold. The price tag for entry-level PLCs is comparable to a COTS server that can host several vPLC instances, being kept out of the factory floor or industrial environment. Distributed I/O on converged Ethernet also provides cost-effective performance and reliability benefits, as communications between different vPLC instances can take place across the convergent fabric or even locally, if co-located on the same host, with SDN allowing for flexible creation of communications channels, for differentiated requirements. Moreover, I/O modules – the components with highest failure rate in PLCs – can be easily and quickly replaced, in case of failure.

Particularly, the potential advantages of the vPLC in terms of reliability, safety and security are considerable, as it can take advantage of datacenter-like redundant power, computing and communications resources. Other benefits are also envisioned, namely:

- Hypervisors allow for migration of virtualized ICS components, as well as instance cloning for pre-deployment tests;
- PLC watchdogs and system-level debugging and tracing mechanisms can be implemented at the hypervisor level, which is able to oversee and control the vPLC partition behavior;
- vPLCs benefit from partitioning isolation, with VMs being easy to restore in a fresh state in case of tampering or other malicious activity;
- SDN-managed isolated I/O paths ease the implementation of flexible, on demand, protection mechanisms at the I/O level (as shown in Section 3) also paving the way for the introduction of NFV components at the ICS level.

Overall, these benefits suggest that virtualizing a PLC could be feasible even for a single instance per device, using Industrial-grade Single Board Computers, instead of COTS servers.

5. Conclusion

This paper discussed the implications of the progressive introduction of virtualization technologies in ICS, with a special focus on security and reliability aspects. The virtualization of both network and computing virtualization was analyzed from an ICS-centric standpoint, covering recent developments as well as proposing new use cases and approaches to improve network and systems security.

Starting with an overview of network virtualization technologies such as SDN and NFV and their application within ICS and distributed IACS, the paper next addressed the issue of using hypervisor technologies for real-time workloads. In this latter perspective, a virtual PLC (vPLC) architecture was discussed, which transcends the simple virtualization of the PLC device, constituting an integrated approach where the device merges with the infrastructure, in a seamless way. The vPLC takes advantage of network and computing virtualization technologies to propose a converged approach for plan-wide consolidation of the ICS infrastructure, with performance, cost and security benefits. This proposal is presently under development by a team that includes the authors of this paper.

Acknowledgements

This work was partially funded by the ATENA H2020 Project (H2020-DS-2015-1 Project 700581).

References

Alves-Foss, J., Harrison, W., Oman, P., & Taylor, C. (2006). The MILS Architecture for High Assurance Embedded Systems. International Journal of Embedded Systems, 2(3-4), 239–247.

Baumann, C., Bormer, T., Blasum, H., and Tverdyshev, S. (2011). Proving Memory Separation in a Microkernel by Code Level Verification. In proc of the 14th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (pp. 25–32).

Beck, M., and Kagan, M. (2011). Performance evaluation of the RDMA over Ethernet standard in enterprise data center infrastructure. In proc of 3rd Workshop on Data Center - Convergent and Virtual Ethernet Switching.

Cahn, A., Hoyos, J., Hulse, M. and Keller, E. (2013) Software-Defined Energy Communication Networks: From Substation Automation to Future Smart Grids. In proc of IEEE SmartGridComm 2013 Symposium - Smart Grid Services and Management Models.

Chavez, A.R., Hamlet, J., Lee, E., Martin, M. and Stout, W. (2015) Network Randomization and Dynamic Defense for Critical Infrastructure Systems. California, USA: Sandia National Laboratories.

Chiosi, M., et al. (2012). Network Functions Virtualization – An Introduction, Benefits, Enablers, Challenges & Call for Action. Issue 1. ETSI White Paper. October 2012. Retrieved February 2016 from http://portal.etsi.org/NFV/NFV_White_Paper.pdf.

Codesys GmbH. CODESYS Control RT: Real-time SoftPLC under Windows. Retrieved March 2016 from https://www.codesys.com/products/codesys-runtime/control-rte.html.

Crespo, A., Ripoll, I., and Masmano, M. (2010). Partitioned Embedded Architecture Based on Hypervisor: The XtratuM Approach. In Proc. of European Dependable Computing Conference (EDCC).

Cruz, T., Simões, P., Monteiro, E., Bastos, F., and Laranjeira, A. (2015). Cooperative security management for broadband network environments. Security and Communication Networks, 8(18), 3953-3977.

Didier, P., and et al., F. M. (2011). Converged Plantwide Ethernet (CPwE) Design and Implementation Guide.

Dong X., Lin H., Tan R., Iyer R. and Kalbarczyk Z. (2015), Software-Defined Networking for Smart Grid Resilience: Opportunities and Challenges, Proc. of 1st ACM Cyber-Physical System Security Workshop (CPSS'15), Singapore, 2015.

ECI Telecom (2015) LightSEC NFV-based Cyber Security Solution for Utilities, Retrieved February 2016 from: http://www.ecitele.com/media/1225/eci_lightsec_nfv_brochure-utilities.pdf

Ersue, M. (2013). ETSI NFV Management and Orchestration - An Overview", Presentation at the IETF #88 Meeting, Vancouver, Canada, November 3-8, 2013. Retrieved February 2016 from: http://www.ietf.org/proceedings/88/slides/slides-88-opsawg-6.pdf.

Fuchs, C. (2012). The Evolution of Avionics Networks From ARINC 429 to AFDX. In Proc. of Innovative Internet Technologies and Mobile Communications and Aerospace Networks (Vol. 65, pp. 65–76).

Galup S. et al. (2009) 'An overview of IT service management, Communications of the ACM, 52(5), pp. 124-127, 2009, doi: 10.1145/1506409.1506439.

García-Valls, M., Cucinotta, T., and Lu, C. (2014). Challenges in real-time virtualization and predictable cloud computing. Journal of Systems Architecture, 60(9), 726–740.

Genge, B., Haller, P., Beres, A., Sándor, H. and Kiss, I. (2016) Securing Cyber-Physical Systems. In Using Software-Defined Networking to Mitigate Cyberattacks, pp. 305-329, 2016, Taylor & Francis Group.

IEEE, Time-Sensitive Networking Task Group. Retrieved February 2016 from: http://www.ieee802.org/1/ pages/tsn.html.

Igure, V.M.; Laughter, S.A. and Williams R.D. (2006) Security issues in SCADA networks, Computers; Security, Volume 25, Issue 7, Pages 498-506, 2006.

Intel Corporation. (2013). Reducing Cost and Complexity with Industrial System Consolidation. Retrieved March 2016 from: http://www.intel.com/content/www/us/en/industrial-automation/reducing-cost-complexity-industrial

IntervalZero. (2010). A Soft-Control Architecture: Breakthrough in Hard Real-Time Design for complex Systems. Retrieved from http://intervalzero.com/assets/wp_softControl.pdf

Irfan, N. and Mahmud, A. (2015) A Novel Secure SDN/LTE based Architecture for Smart Grid Security. In proc of IEEE International Conference on Computer and Information Technology.

ISA-99.00.01 (2007) Security for Industrial Automation and Control Systems - Part 1: Terminology, Concepts, and Models, American National Standard.

ISaGRAF. ISaGRAF Overview. Retrieved February 2016 from: http://www.isagraf.com.

Kang, D. et al., (2011) Proposal strategies of key management for data encryption in SCADA network of electric power systems, Int. Journal of Electrical Power & Energy Sys., Vol. 33, Iss. 9, Nov. 2011.

Kean, L. (2010). Microcontroller to Intel Architecture Conversion: PLC Using Intel Atom Processor.

Kreutz, D., Ramos, F., Verissimo, P., Rothenberg, C., Azodolmolky, S. and Uhlig, S. (2014). Software Defined Networking: A Comprehensive Survey. Proc. IEEE, 103(1), pp.14-76.

Krutz, R. L. (2006) Securing Scada Systems, USA: Wiley Publishing, Inc., 2006.

L. O'Murchu, N. Falliere (2011) W32.Stuxnet dossier, Symantec White Paper, February 2011.

Machii, W., Kato, I., Koike, M., Matta, M., Aoyama, T., Naruoka, H., Koshima I. and Hashimoto, Y. (2015) Dynamic Zoning Based on Situational Activitie for ICS Security. In IEEE 978-1-4799-7862-5/15.

ONF (2012). OpenFlow Switch Specification, version 1.3.0 (Wire Protocol 0x04), Open Networking Foundation,

Proença, J., Cruz, T., Monteiro, E., and Simões, P. (2015). How to use Software-Defined Networking to Improve Security–a Survey. In proc of the 14th European Conference on Cyber Warfare and Security 2015 (pp. 220).

RAD Data Communications Ltd. (2015) Megaplex-4 D-NFV Virtualization Module, Retrieved February 2016 from: http://www.rad.com/Media/34173_D-NFV.pdf.

Siemens AG. Jailhouse Partitioning Hypervisor. Retrieved March 2016 from: https://github.com/siemens/jailhouse

Silva, E.G., Knob, L., Wickboldt, J., Gaspary, L., Granville, L. and Schaeffer-Filho, A. (2015) Capitalizing on SDN-based SCADA systems: an anti-eavesdropping case-study. In IFIP 978-3-901882-76-0.

Triangle MicroWorks, Inc (2002) DNP3 Overview, Raleigh, North Carolina, Retrieved February 2016 from: http://www.trianglemicroworks.com/documents/DNP3_Overview.pdf.

X-HYP Project. X-HYP Project. Retrieved February 2016 from: http://x-hyp.org.

Zhang, W., Wood, T., Ramakrishnan, K., and Hwang, J. (2014). Smartswitch: Blurring the line between network infrastructure \and cloud applications. In Proc. of 6th USENIX Work. on Hot Topics in Cloud Computing.

Heuristic and Proactive IAT/EAT-Based Detection Module of Unknown Malware

Baptiste David, Eric Filiol, Kevin Gallienne and Olivier Ferrand
ESIEA - Laboratoire de Cryptologie et de Virologie Opérationnelles, France
eric.filiol@esiea.fr

Abstract: In the context of the DAVFI National research Project, we have designed a heuristic algorithm which is able to detect both known and unknown malware (binary executable files) very accurately while requiring a very limited number of updates. Our method is based on an original approach which mostly use supervised learning algorithms when considering vectors built with relevant information extracted from binary executable files (Import Address Table [IAT] and Export Address Table [EAT]). Our heuristic module is designed to detect both unknown malware and known malware. The overall performance gives a true positive rate of at least 96 % and false positive rate of less than 4 % (with respect to truly unknown malware). Moreover, our module is chained with other modules (classical black listing techniques, combinatorial detection module…) and the overall performances yields a true positive rate of 99 % with a false positive rate which tends towards 0.

Keywords: malware, antivirus, targeted attack, heuristics

1. Introduction

The DAVFI project (standing for *French and International Antiviral Demonstrator*) was a 2-year project partially funded by the French Government (DAVFI, 2012). This project aimed at designing, implementing and testing a proof-of-concept for a new generation, sovereign, open antivirus software. Based on a strongly multithreaded architecture, the final demonstrator is made of several modules chained into two main entities: a resident kernel notification driver and an antiviral analysis service. The latter embeds two analysis streams: one for binaries and executable files the other to process documents (and malware documents) specifically (Dechaux & Filiol, 2015).

To summarize, this analysis and detection chain combines structural analysis, dynamic white-listing and black-listing modules (modules 1.1, 1.2 and 1.3) with a signature-based detection module (optimized SEClamav module 5.1) and our heuristic, proactive detection module (module 5.2) which is presented in the present paper. DAVFI resources (source code, data, technical documentation...) will made as open and free as possible in 2016.

The major constraint in DAVFI's specifications was to design an antivirus engine which outperforms the best existing commercial AV software, especially regarding the proactive detection of unknown malware (that is to say before updating the engine malware databases).

To detect unknown malware (or at least malware that are unknown from the antivirus database), heuristic methods or more generally statistical approaches are the most promising research trends nowadays. However innovative detection algorithms cannot be included in antivirus software due to performances requirements. Among them, we can talk about a relatively high false positive rate, the time response for a given sample or memory limits. Having a too high false positive rate may be a critical issue regarding executable files which are essential for the operating system kernel, for instance. Reducing the risk of false positive detection by limiting the scope of efficient heuristic methods is still possible but it does not constitute a realistic solution.

Most of commercial AV products rely on signature-based detection (opcodes, control flow graph…). Even when heuristics are supposedly used, they do not capture and synthetize enough information to be able to detect unknown malware accurately and proactively. This implies that frequent and prior updates must be performed. May their analysis techniques be fully static of dynamic (using sandboxing or virtual machines), commercial AVs do not capture what defines malware compared to benign files: their intrinsic actions. In this paper, we describe how effectively synthetize these actions and the difference between malware and non-malicious files. We extract and analyze two tables that are present in executable files: the Import Address Table [IAT] and Export Address Table [EAT]. These tables summarize the different interactions of the executable with the operating system. We show how this information, once it has been extracted, can be used in supervised learning to provide an effective detection algorithm which have proven to be very accurate and proactive with respect to unknown malware detection.

The paper is organized as follows. Section 2 presents the technical information we extract to build the training sets which be used by our detection algorithm and why we focus on these data. Section 3 then presents our supervised detection algorithm: building of the training model and the detection algorithm itself. We also discuss in this part the mathematical validity of this algorithm. Section 4 presents the results we achieve and discuss the detection performances. In Section 5, we will conclude and present future work to develop this heuristic module further.

In this paper, the terms 'benign file' and 'goodware' describes indifferently the same reality.

2. Technical background: Import Address Table [IAT] and Export Address Table [EAT]

Because most of the malware are targeting the Windows system, our algorithm is mostly designed for this operating system family. However our approach has been similarly extended and applied to UNIX systems in the same way (up to the technical differences between ELF executables and MZ-PE executables). Even if we implemented our algorithm to be able to detect UNIX malware specifically as well, without loss of generality we will not present it in this paper since it would be a recurrence of what has been made for Windows.

2.1 Introduction to IAT and EAT

Any executable file contains a lot of information in the MZ-PE header (Microsoft, 2013) but some information can be considered more relevant than the others. Tables like *Import Address Table* (IAT) and *Export Address Table* (EAT) are, in our case, enough to describe what a program should do or is supposed to do. The IAT is a list of functions required from the operating system by the program. Technically there are two possibilities of importing functions on Windows. The firt one is made explicitly through the IAT during the loading phase of the process before running and or during the running phase with the use of the *LoadLibrary* and *GetProcAddress* functions (MSDN, 2015a & 2015b). The second possibility is used by a lot of malware to hide their real functionalities by loading them without referencing them in their IAT. Nonetheless, the functions are used to load libraries and to retrieve functions during runtime and therefore constitute some unavoidable *points of passage* which can be referenced. In most of the cases, malware or packers have enough significant IAT to be detectable.

All executable files need an IAT. Without IAT - if this one is empty - it would mean that the targeted program would have no interaction with the operation system. In other words, it is not able to display text or any information at screen, it is not able to access any file on the system and it cannot allocate any segment of memory. Except consuming CPU time - with no result exploitable - it is not supposed to do anything else. Such useless program can be considered as suspicious (since it is suspicious to launch useless programs) or as malware in the most common case. If executable files need IAT, Dynamic Linked Library (Dll) can also provide an EAT. This table describes which functions are exported by a Dll (and which are importable by an executable). Dll generally contains IAT and EAT - except for specific libraries which only export functions or objects. An executable can contain both an IAT and an EAT (the kernel of Windows *ntoskrnl.exe* is a good example). The use of EAT and IAT is a good combination to discriminate most of the libraries since the export and import is quite unique.

However, there are some limits to this system. One lies on the fact that this system only uses and trusts function, executable or library names. If a malware is designed to change every name of function to unknown ones, the system will not be able to give any reliable information any more. In addition, samples which *imitate* IAT and EAT from real goodware (benign files) are able to bypass this type of test. Of course, it is a true limit of our model but, surprisingly, in most operational case, such a situation is not common. Most of the packers which are used on malware provides reliable IAT and EAT based on the executable file packed or on the packer itself (which helps to discriminate which packer is used). This observation is extensible to setup programs which are sort of packers.

To be really accurate in all situations, our detection algorithm has to be chained completed with other tests - based on data and opcodes sections analysis or on file header structural analysis (David & al., 2016). But for real-time analysis, in the situation where our detection algorithm is one antivirus filter which is combined with others, results are enough to achieve a quite acceptable trade-off between efficiency and reliability.

2.2 IAT and EAT extraction

Before we can extract the IAT and EAT, it is necessary to find whether they are present or not. For this purpose it is necessary to analyze the entries of each table in the *DataDirectory* array of the IMAGE_OPTIONAL_HEADER (or IMAGE_OPTIONAL_HEADER64 in x64) structure. These entries (whose type is IMAGE_DATA_DIRECTORY) are *DataDirectory*[IMAGE_DIRECTORY_ENTRY_EXPORT] and *DataDirectory*[IMAGE_DIRECTORY_ENTRY_IMPORT]. For the IAT and EAT to be present, it is necessary that the *VirtualAdress* and *Size* fields in the associated structures are non-zero.

Upon confirmation of the presence of an IAT, it must then be read. Each DLL is stored as a structure of type IMAGE_IMPORT_DESCRIPTOR. From this structure is extracted, at first, the *Name* field that contains the name of the DLL, then the *OriginalFirstThunk* field containing the address where is stored the primary function, the other being stored in sequence. Each function is stored in a structure of type IMAGE_THUNK_DATA, in which the field *AdressOfData* (whose type is IMAGE_IMPORT_BY_NAME) contains

- the hint value (or *Hint* field). This 16-bit value is an index to the loader that can be the ordinal of the imported function (Pietrek, 2001);
- and the function name, if present (*Name* field), i.e. if the function has not been imported by ordinal (see further in Section 2.3). In the case of imports by ordinal only, it is the *Ordinal* field of IMAGE_THUNK_DATA that contains the ordinal of the function (if the most significant bit is equal to 1 then it means that the low 16 bits are the ordinal of the function (Iczelion, 1996)).

After getting the name of the function, a pair dll_name/function_name (function_name is the name of the function or its ordinal otherwise) is formed and stored, and the next function is played until all the functions of the DLL are read, and so on for each imported DLL. On output, a set of pairs *dll_name/function_name* is obtained, which will go through a formatting phase (see Section 2.4).

The format of the EAT, although also representing a DLL and all of its functions, is different from that of the IAT. All of the EAT is contained in a structure of IMAGE_EXPORT_DIRECTORY type. From this structure are obtained the name of the DLL (which may be different in the case of renaming) using the *Name* field, the number of functions contained in the EAT (*NumberOfFunctions* field) and the number of named functions among them (since some functions can be exported by ordinal only) (NumberOfNames field).

Then we recover the functions and their name/ordinal. For the named functions, we just have to read in parallel two arrays whose addresses are *AddressOfNames* and *AddressOfNameOrdinals*: at equal index, one contains the name of a function, and the other, its ordinal. For non-named functions, we must then retain all ordinals of named functions and then recover in the table with address *AddressOfFunctions* - which is indexed according to the ordinals of the functions it contains - all the functions whose ordinal has not been retained. After obtaining the set of functions/ordinals, in a similar way to that for the IAT, a set of pairs dll_name/function_name} is formed and then formatted (Section 2.4).

2.3 Miscellanous data

Let us now detail a few technical points that are interesting to understand IAT and EAT in depth. Microsoft's documentation (Microsoft, 2013b) explains how to export functions by ordinal in a DLL: ordinals inside a DLL MUST be from 1 to N, where N is the number of functions exported by the DLL. This is interesting and leads us to think that maybe some malicious files do not respect this rule. To go further, it is likely that this also applies to the *hint* of functions, although no documentation about it could be found. However, the analysis of a few Windows system DLL export tables like *kernel32.dll* and *user32.dll* shows that they comply to this rule. After conducting tests on malicious files and benign files, it turns out that only one `healthy' file (*sptd.sys*, a driver from alcohol120%) does not follow this rule, while a number of malicious files do the same.

2.4 Generation of IAT and EAT vectors

After getting all the *dll_name/function_name* pairs from a file, two vectors are created (one for the IAT and one for the EAT). These vectors will the base object for our detection algorithm. In order to generate those vectors, we must build a database containing all the known pairs. A unique ID is associated to each unique pair. This database is populated by a base set of files with a known classification (malicious or benign). The population process is the following:

- EAT and IAT pairs are extracted from a file.
- For each pair, a unique ID is constructed. This ID is a 64-bit number with the high 20 bits representing the DLL and the remaining 44 bits representing the function.
- For the DLL ID: if the DLL is known, its ID is used. In the other case, a new ID is used, corresponding to the number of currently known DLLs (the first is 0).
- The function ID follows the same process with known functions.

This population process is only executed manually whenever we want to update the database; it is not run during file analysis. The two vectors are created according to this database. For each pairs, its ID is recovered from the database. If it does not exist in the database, the pair is discarded. All the 64-bit numbers are then sorted and stored in a file.

3. The detection algorithm

In this section we are now presenting our supervised detection algorithm which works on the vectors built with the data extracted and presented in the previous section. Usually (Maloof, 2006; Rajaraman & Ullman, 2011; Williams & Simoff, 2013) the database of known samples (training set) must be built before writing the detection algorithm, as far as supervised algorithms are concerned. Such a procedure is led by the knowledge and the learning of what to detect (malware) and what not to detect (benign files). So the training set contains two subsets summarizing the essence of what malware and benign files really are.

3.1 How to build the algorithm

Our solution is quite different. Indeed, if we know beforehand which data to use to perform detection, we did not know how to build the database to make it reliable and accurate enough for our algorithm. Which data to select among a set of millions of malware samples and of benign files, in order to get a representative picture of what a malware is (or is not) for the algorithm, is a complex problem in itself. Our approach has privileged the operational point of view. We have designed the algorithm as formal as possible and we have applied it on sets of malware and on a set of benign files to allow it to learn by itself, building the database after the creation of the algorithm. In other words, the algorithm is designed to use a minimal database of malware and of benign files at the beginning and this one is able to perform minimal detection helping to develop the database with samples undetected to improve results. We thus consider an iterative learning process, somehow similar to boosting procedure (Hastie & al. 2009; Williams & Simoff, 2006).

Such an approach privileges experimental results and design of algorithms to detect unknown malware. Indeed, the algorithm uses subsets of malware samples which are the most representative of their families. Derivatives and parts of known malware (or variants) can be recognized since they have been learned previously. `Unknown' malware use most of the time *old fashion* technologies, same base behaviors and hence our algorithm is able to detect a lot of them with such a design and approach. For sake of clarity, the description of our algorithm starts with building the detection databases (training set). To help the reader, we suppose in this part that we (already) have a known detection algorithm which is presented right after in the paper (section 3.3).

3.2 Building the detection database (training set)

The heuristic algorithm we have designed uses a database of knowledge to help it to make decisions. Of course, algorithm databases are built with the two different types of files it is supposed to process and decide on: benign files and malware. The use of a combination of samples from *malware* and *goodware* gives the best results since they are suitably chosen. The way the database is built is the key step of our heuristic algorithm, since it affects directly the results we obtained. However, we must stress on the fact that we would obtain the same results for different malware/benign files subsets, as long as those sets are representative enough of their respective family. Somehow, this step can be seen as a probabilistic algorithm.

From a simple observation, more than the number of samples we could set in the database, the diversity of samples helps better to get the widest possible spectrum of detection. Smaller and more diverse the database is, faster and better are the results given. Indeed, if the database is too big, searching inside will be too much time-consuming, thus resulting in the impossibility to use it in real time. Only the most representative malware of a family must be included in the database (and similarly for the benign files).

To build the database, we need to select the most relevant files to store. First, we need a detection function which is the one used by our algorithm. At the beginning, the database used by this function is composed only with a small set of malware arbitrarily selected (denoted M) to be representative of the family we want to include. Such a detection function can be defined as follows. From any sample S we want to detect, we have a prior detection function D_M which is of the form

$$D_M(S) = \begin{cases} 0 \; if \; S \; is \; a \; goodware \\ 1 \; if \; S \; is \; a \; malware \end{cases}$$

The function D_M does not need to exhibit huge and optimal detection performances. So a known and initial malware (respectively benign file) sample set is enough to initiate the process.

Data: A set of files to analyze S_f (which has n files) and a maximal error detection rate ϵ.
Result: A set of file M containing the database
while $\frac{|S_f|}{n} < \epsilon$ **do**
  for $\{s\} \in S_f$ **do**
    if $D_M(s) == 1$ **then**
      $S_d \leftarrow \{s\}$;
    end
    else
      $S_{ud} \leftarrow \{s\}$;
    end
  end
  $M = M \cup S_{ud}$;
  if $|S_d| == 0$ **then**
    break;
  end
  $S_f \leftarrow S_{ud}$;
end

Figure 1: Training set building algorithm

To expand the databases (malware and goodware), Algorithm 1 (Figure 1) is used. This approach is more or less similar to boosting methods such as Ada-Boost (Freund & al., 1997; Hastie & al., 2009).

Algorithm 1 also enables to control the error detection rate ε for a given malware family (with $\varepsilon \in [0, 1]$). Indeed, if ε chosen is too small, the algorithm can include all the files from S_f (see Figure 3). Of course, the representativeness of files in S_f is a key point to use the algorithm. Working with several different samples of the same family is, most of the time, the best approach. Another possibility of control is to use the rate of detected files such as $\frac{S_d}{S_{ud}} < \varepsilon$ with ε close to zero.

The building of database is performed family per family (of malware). It is possible to make it faster mixing multiple relevant samples from different families in one set. For example, to build the goodware database, one can chose files among those coming from C:\Windows. In fact, the initial choice of incoming files defines the relevance and the diversity of the database. Starting from a small set of these files, we launch Algorithm 1 on the remaining files until we have got enough file detected by the database created on the fly.

One key advantage of this principle lies in the fact that we can increase the size of the database in the future without prior knowledge of a malware family. At the first time we created the database, if the diversity of malware families was enough good, it is possible to include new samples of malware without knowing its type/family. In fact, malware share strong IAT and EAT correspondences and similarities with many other families, in most of the cases. It means that malware can be detected by the database previously built even if we never included any sample from its family. In other words, we can use this property to increase the size of the database by adding undetected malware coming from different families into the current database. Taking a file defined as malware (which could be given by any trusted source or by a prior manual analysis), if this one is

not detected by our algorithm, we can include it in our database in order to improve the detection of its family. It is a simple way to improve the accuracy of the detection.

3.3 The detection procedure: The K-nn algorithm

Once the structural analysis is achieved and the database (training set) has been built, then the detection tests occur by using the IAT and EAT vectors which have previously generated. This is the second part our module is in charge of, and which aims to be able to decide the nature of a file.

Detection tests are split into two sets: the IAT comparison test and the EAT comparison test. The principle of those tests is: the unknown file's IAT (or EAT) is compared to each element of the base of benign files and to each element of the base of malicious files. The k files that are closest to the unknown file are kept with their respective label (malware or goodware). A decision is then made based on these k files to decide which label to give to the file under analysis. This test thus uses the method of k-Nearest Neighbours (Hastie & al., 2009; Williams & al., 2006), which has been modified for the occasion. In both cases, the input consists of the k closest training examples in the feature space.

3.3.1 Vector format limits

While this format allows an optimized storage of the IAT/EAT, it faces several constraints that limit its use. The first constraint is a space constraint, which actually is not an intractable problem. Our encoding limits to 2^{20} possible DLLs and to 2^{44} functions per DLL. Today, this is more than enough, but we must keep in mind that this limit exists, and could be a problem in a (very far) future.

The second constraint lies in the fact that our vectors do not have a fixed length. It is a problem if we want to use standard distance functions, like the Euclidean distance. We could have used a similar vector format in which each possible couple was given a 0 or 1 number depending on whether it was present in the file or not. But the length would have been around 10^6 (about the current size of the database) instead of around 10^3 (for large files) with the current format. It would have a bad impact on the performances of real-time analysis, and hence it would have increased the time of analysis by a too high factor. In order to optimize the computation time, all the vectors in the bases and generated during analysis are sorted.

3.3.2 The similarity measure

In order to determine the nearest neighbours, we need a function to compare two IAT/EAT vectors of different sizes. The format prevents the use of standard distances (because to use a standard distance, the IAT/EAT vectors should have the same size, i.e. always the same number of imported/exported functions in each file, which is quite never the case). It was therefore necessary to find a function fulfilling this role and to apply it our format. Let us adopt a few notations:

- An IAT/EAT vector of size n is written as $\sigma = \sigma_1\sigma_2\ldots\sigma_n$ where $\sigma_i \in \{0, 1\}^{64}$ (64-bit integers).
- I: E, F $\rightarrow$ {0, 1} such as $\forall x \in E$, $I_F(x) = 0$ if $x \in F$ and 1 otherwise.
- If v is an IAT/EAT vector, $E_v = \{\sigma_i\}$ (this notation describes the fact that vectors are implemented as lists of 64-bit integers).

The function used which defines a degree of similarity between IAT or EAT vectors is

$$\forall a \in \Sigma_U, \forall b \in \Sigma_U, f(a,b) = \frac{1}{|a|+|b|}\left(\sum_{i=1}^{|a|} I_{E_b}(a_i) + \sum_{j=1}^{|b|} I_{E_a}(b_j)\right)$$

It is easy to prove that this function satisfies the separation, the symmetry and the coincidence axioms as a similarity measure must.

3.3.3 The decision algorithm

The detection algorithm which is used to decide the nature of a file (malware or benign) is given by Algorithm 2 (Figure 2). It is composed of two parts in order first to reflect the importance of similarity optimally and second to eliminate some neighbours who are there only due to the lack of data.

```
Data: A vector X representing a file to analyze, a malware vector base B_M and a
      benign vector base B_B.
Result: A Boolean value indicating whether the file is malicious or not.
i = 0;
for {b} ∈ B_B do
    d = f(X, b);
    if d == 0 then
        return false;
    else
        neighbors[i] += (d, benign);
        i++;
    end
end
for {m} ∈ B_M do
    d = f(X, m);
    if d == 0 then
        return false;
    else
        neighbors[i] += (d, malicious);
        i++;
    end
end
if MaxNeighbors(neighbors) == malicious then
    return true;
else
    return false;
end
```

Figure 2: Algorithm 2 used to classify a file

The first part consists in filtering the set of neighbours that the k-NN algorithm returns to refine the best decision based on the neighbours that are really close. For this, a threshold is set (50 % for now) and only neighbours with a higher degree of similarity (i.e. that the function f returns a value less than 0.5) are kept. Then classical decision is applied to this new set: the file is considered closer to the base with the most representative among the neighbours.

The second part is used in the case when an equal number of representatives in each base, is returned (situation of indecision). All the neighbours are again considered, and again the file is considered closer to the base with the most representatives among the neighbours. If k is odd, it helps to avoid indecision (majority decision rule). It was therefore decided that all k are used odd in order not to fall in the case of indecision.

4. Detection and performances results

In order to test and to tune up our algorithm, we have defined many tests. On the one hand, we have tested the modification of the number of neighbours' parameter in the k-NN algorithm. This test is made in order to observe for how many neighbours the test is the most efficient. Then, on the other hand, we performed tests on databases to measure results of the algorithm. Of course, the detection algorithm is used with the most efficient number of neighbours obtained in the first test.

Increasing the number of neighbours more than 9 does not change the results significantly. In fact, keeping the number of neighbours as minimal as possible is a better choice since it increases the response time of the algorithm - a key point when we used it in real-time detection conditions. The results about this test are displayed in Figure 3.

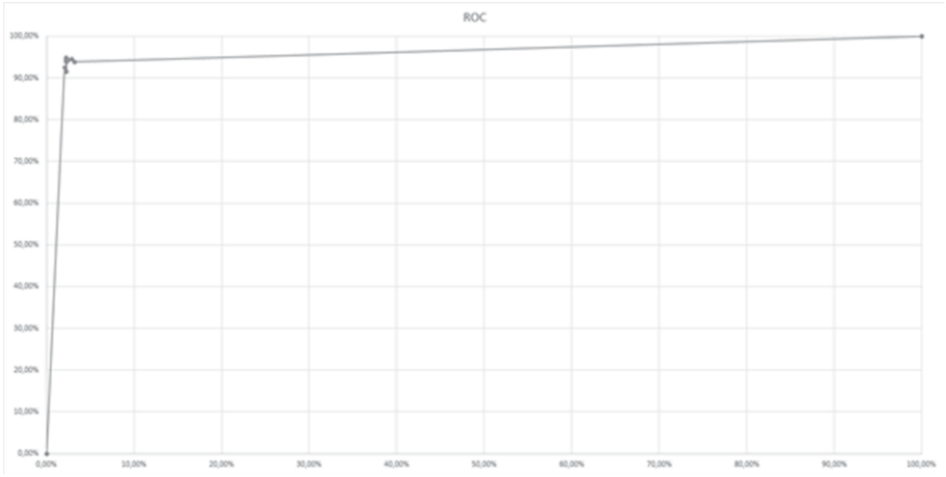


Figure 3: ROC summarizing the detection algorithm performance

For the final test, we have put the algorithm to the proof with two databases. One is composed of 10,000 malware (extracted from different families and unknown from our databases) and one composed of goodware composed of executable files extracted from Microsoft Windows operating system (around 131,000 files). The results are given in Table 1.

These results show that the algorithm is quite efficient to detect similarities between different executable files. Nonetheless, it is not enough to use it for detection in real time only since the rate of false positive detection is too high to be acceptable. To prevent such a case, our algorithm in module 5.2 is chained with other modules performing structural analysis (David & al., 2016), white-listing and black listing filtering (see Figures 1 and 2). This is the most efficient approach since we succeeded in making the residual false positive rate tends towards 0.

Table 1: Results

	Malware database	Benign files database
Detected as malware	95,028 %	4,972 %
Detected as benign file	2,053 %	97,947 %

Once we have validated our algorithm and confirmed its performances experimentally, we had the opportunity to test it in operational and real-life conditions. A users committee was present in the DAVFI/OpenDAVFI project. The aim was to involve end users, to have their operational feedback regarding antivirus software and to make them test a few modules in real-life conditions. Moreover, they feed us with unknown malware (usually manually detected in their CERT during the very first hours of the attack), most of them being not detected by commercial AV software (we use the VirusTotal website for checking this point). Most of the samples provided related to targeted attacks.

The first experiment concerned a blind detection on a set of unknown malware coming from a French governmental entity. The detection results (true positive rate, false positive rate) comply with those presented previously. The second experiments involved a major French bank which provided a number of targeted malware unknown to VirusTotal. Our algorithm systematically succeeded in deciding them as malicious.

5. Conclusion

In this paper we have presented a supervised detection algorithm working on data extracted from the IAT and EAT of binary executable files (Windows and Unices). These particular pieces of information do not only describe the executable in a static way more precisely (use of far more complex and rich signatures) but also they capture the information related to program behaviors. The overall performances which we have achieved show that is possible to detect unknown malware proactively and accurately. This yields enhanced detection capabilities while requiring far less database update. Beyond the experimental analysis, operational testing of our algorithm has been performed on malware coming from the real world in real conditions. The results which have been observed fully satisfy the operational constraints and specifications of the project.

Future work will address the combinatorial modelling and processing of information contained in IAT/EAT. While we have considered mostly statistical aspects, it is possible to have a far more precise processing of this information when using combinatorial structures to synthetize the concept of behaviors and hence base a more accurate detection on the dynamical information contained in the code. We also intend to extend the information used for detection. The study of data section or opcodes sections is a possible option in order to increase the number of detection criteria. These sections can provide correlations with the features we already consider.

References

David, B., Filiol, E. and Gallienne, K. (2016). *Structural analysis of binary executable headers for malware detection optimization*. To appear in Journal in Computer Virology and Hacking Techniques.

DAVFI Project Website (2012 - 2014), http://www.davfi.fr/index_en.html.

Dechaux, J. and Filiol, E. (2015). Proactive defense against malicious documents. Formalization, implementation and case studies. To appear in Security Special Issue, Roy Park editor, Journal in Computer Virology and Hacking Techniques.

Freund, Y. and Schapire, R. E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of computer and Systems, 55:119:139.

Hastie, T., Tibshirani, R. and Friedman, S. (2009). *The elements of statistical learning: Data mining, inference and prediction*. Springer Verlag

Iczelion (1996). *Win32 assembly. Tutorial 6: Import table*. http://win32assembly.programminghorizon.com/pe-tut6.html

Maloof, M. A. (2006). *Machine learning and Data mining for computer security*. Springer Verlag.

Microsoft MSDN (2013). *Microsoft PE and COFF Specification*. http://msdn.microsoft.com/en-us/library/gg463119.aspx

Microsoft MSDN (2013). *Exporting from a DLL using DEF files*. http://msdn.microsoft.com/fr-fr/library/d91k01sh.aspx

Microsoft MSDN (2015). *Loadlibrary function*. https//msdn.microsoft.com/en-us/library/windows/desktop/ms684175(v=vs.85).aspx

Microsoft MSDN (2015). *Getprocaddress function*. https//msdn.microsoft.com/en-us/library/windows/desktop/ms683212(v=vs.85).aspx [website]

Pietrek, M. (2002). An in-depth look into the Win32 portable executable file format, part 2. http://msdn.microsoft.con/en-us/magazine/cc301808.aspx

Rajaraman, A. and Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge Univeristy Press.

Williams, G. J. and Simoff, S. J. (2006). *Data mining - Theory, methodology, techniques and Applications*. 2nd edition, Springer Verlag.

Conceptualising Cyber Counterintelligence: Two Tentative Building Blocks

Petrus Duvenage[1], Victor Jaquire[2] and Sebastian von Solms[1]
[1]Centre for Cyber Security, University of Johannesburg, South Africa
[2]Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa
duvenage@live.co.za
jaquire@gmail.com
basievs@uj.ac.za

Abstract: Several escalating trends are affirming the centrality of Cyber Counterintelligence (CCI) in effectively addressing advanced cyber threats of today and tomorrow. Yet, in comparison with the burgeoning academic and commercial literature on the related field of Cyber Threat Intelligence (CTI), CCI remains vastly unexplored. Outside the circles of governments' security apparatus, some large corporates and niche vendors that offer such specialised services, CCI is still obscure. While interest is gradually growing in CCI, this academic discipline is very young and largely uncharted. Leveraging off previous research by the aforementioned authors, this paper advances two further building blocks to contribute towards constructing this emerging discipline. Building block 1 comprises a distinction between CCI and CTI. Such a distinction is necessary for clarity and has the advantages of allowing CCI to benefit from the extensive research work done in the CTI field. Building block 2 consists of a multi-layered framework that explicates the different levels on which CCI functions, namely the strategic, operational and tactical functional levels. This framework progresses building block 1. While these functional levels have been described extensively in CTI literature, no such CCI-specific application could be found in literature within the public domain. Since it expounds CCI on the various levels that it functions, the framework contributes to a more nuanced academic conceptualisation of this discipline of CCI. On a practical level, the framework could serve as a notional guide for performing actual CCI work more effectively. The article concludes by reiterating the importance of CCI in addressing advanced threats and suggesting areas for further research.

Keywords: cyber counterintelligence, cyber threat intelligence, offensive cybersecurity, cyber counterintelligence levels, cyber counterintelligence maturity

1. Introduction

In what has become a recurring theme in recent years, industry threat reports for 2015 to 2016 highlighted the escalating damage caused and threats posed by cyber actors of increasing sophistication (Kaspersky 2015, McAfee 2015, Crowdstrike 2016). This trend is accelerating despite a continuing increase in global spend on cyber security. In recent years, vendors have been pushing particularly Cyber Threat Intelligence (CTI) as a critical part of the 'solution' and it has evolved to one of the fastest growing cyber security sectors. The $1, 02 billion global spend on CTI in 2015, for example, represents a 129% increase compared to 2011 (Statista 2016, Info-security Magazine 2015). Further attesting to threat intelligence's rising prominence is the escalation in Google search results from a mere 18 700 in 2011 to 381 000 in February 2016 (Chismon & Ruks 2015, Google 2016).

As matters currently stand, the CTI market buzz and spending of resources have not by any measure translated in a corresponding mitigation of advanced threats – nor is it likely to do so in the near future. There are various reasons for this rather gloomy prognosis of which two will be highlighted in this paper.

The first reason is that a significant portion of products and services and that are marketed as CTI is not intelligence at all. They are mere re-labelled data feeds or anti-virus packages. Of course products of this nature have a role, but they are wholly insufficient against higher-end threats. It can rightly be argued that sound CTI as part of an effective cyber-security approach would be effective in addressing advanced threats. This is indeed the case, but only partially. CTI employed as part of an effective cyber-security approach will address a substantial portion of cyber threats. It will, however, not be effective against those high-end threats that should top our concern. For CTI to be effective against these threats, it needs to be embedded in counterintelligence (CI).

The second reason for CTI not delivering on expectations is that CI is simply not being embraced. Organisations with significant cyber assets are too slow to realise that we are faced with CI challenges rather than cyber security problems. Perhaps, we are still too attached to outdated, neat tables linking specific cyber actor types

to certain methods and aims. In reality, the distinction between what was conventionally labelled as state-sponsored Advanced Persistent Threats (APTs) and the actions of other actors is blurring fast. In its 2015 Global Threat Report, Crowdstrike (2016) states, for example, that "the primary motivation behind global cyber activity has now shifted from disparate activities carried out by individuals, groups and criminal gangs pursuing short-term financial gain, to skilled adversaries driven by broader agendas." The cyber criminals' aim, asserts PwC (2016), currently "goes beyond targeting financial information to include a company's 'crown jewels' – customer data and intellectual property information, the loss of which can bring down an entire business." Various types of threat actors can and do cooperate (INSA 2011). The tradecraft, activities and even aims of various classes of threat actors in cyber space are often difficult to separate and reflect high skill levels in intelligence and counterintelligence (Moyo 2015). For state and non-state actors (such as criminal groups, some corporate entities) multi-vectored espionage (e.g. human and technical means) has become a precursor to extensive breaches. The addressing of such threats is CCI's signature role.

While CI/CCI awareness within board rooms appears to be growing, these concepts are far less known than CTI (cf. SpearTip 2015, The Economist 2015). Moreover, the symbiotic relationship that should exist between CCI and CTI is seldom addressed. Therefore academia has a crucial role in conceptualising CCI clearly. This paper proposes two further building blocks that could aid in conceptualising this discipline. Firstly, CCI is distinguished from CTI and the relationship between these constructs examined. Secondly, a multi-layered framework is submitted to explicate the different levels on which CCI functions. Notionally and practically, this multi-levelled examination provides clarity on what CCI is, what it does and what its relation with CTI is.

It needs to be emphasised that this paper builds on previous articles that defined various CCI concepts, positioned CCI as part of multi-disciplinary CI, detailed CCI's defensive-offensive modes and advanced a process model (Duvenage & von Solms 2015; Duvenage, von Solms & Corregedor 2015). While some aspects of previous work are concisely recapitulated (per Section 2.2), the latter is highly selective and could not address all aspect necessary for context.

The foregoing introduction highlighted the importance of CCI in addressing current and future cyber threats. Subsequently, the need to further conceptualise CCI was underlined. The next section delineates CCI and CTI by offering definitions and discussing the relationship between the constructs.

2. Conceptual clarification – what are 'threat intelligence' and 'cyber counterintelligence'

As suggested above, CTI and CCI are interrelated yet distinct concepts. Delineating these two constructs is important, since each has a unique and complementary role in ensuring cyber security. Moreover, a clear differentiation would enable CCI to draw on extensive CTI literature in a manner that is academically credible and responsible.

2.1 Defining 'cyber threat intelligence'

The rapid market growth in the market of CTI products and services has been accompanied by a proliferation in terms and definitions. "Threat intelligence", "cyber intelligence", "cyber threat intelligence" are sometimes used interchangeably and sometimes with different connotations (Deloitte 2014, Schoeman 2015, EMC^2 2014, INSA 2013, INSA 2014*a*-*b*, Lee 2014*a*-*b*.). A dissection of all these terms will distract from the paper's main focus and be more confusing than helpful. In the interest of simplicity CTI is henceforth employed in the paper as the umbrella term. Schoeman (2015) rightly states that CTI has evolved in a "catchall term for a vast array of different technologies, methodologies and ideas." Products and services sold under this banner can vary extensively in scope, usability, aims and contents (Chismon & Ruks, 2015). At the one end of the spectrum CTI can be just anti-virus signatures at a much higher cost; while at the other end, it can mean an overarching approach central to an organisation's strategy (Schoeman 2015, Riley 2015).

The term 'threat intelligence' has its roots in the concept 'intelligence' as used with state security apparatus and Intelligence Studies. Depending on context, 'Intelligence' can have several meanings within Intelligence Studies. Intelligence can denote the overarching discipline that comprises Positive Intelligence, Counterintelligence and Covert Action. Sometimes Intelligence is often employed as a shortened reference to Positive Intelligence. The term Intelligence could furthermore refer to the outcome of a process that delivers actionable, analysed information. These meanings and applications thereof in the cyber realm were explored in an earlier article (Duvenage, von Solms & Corregedor, 2015). Suffice it to note here that 'intelligence' used in 'cyber threat

intelligence' – and as henceforth applied in this paper – means actionable, assessed information on a cyber-related hazard to an entity. This is in line with Gartner's defining of threat intelligence as: "evidence-based knowledge, including context, mechanisms, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard" (Schoeman 2015). Deriving intelligence from information and data requires analysis performed by humans. Tools and data feeds cannot by themselves provide threat intelligence (Schoeman 2015). In this regard Lee (2014*a*) states "Intelligence of any type requires analysis. Analysis is performed by humans. Automation, analytics and various tools can drastically increase the effectiveness of analysts but there must always be analysts involved in the process." In summary it can thus be said that data is processed and refined to produce information. Information is in turn analysed and presented in a format that is actionable and constitutes intelligence. In the case of CTI, this is intelligence produced on cyber-related hazards.

Ideally CTI should provide intelligence on a full spectrum of adversarial action in the cyber sphere from decision to execution. INSA (2013) provides the following breakdown of these actions and what cyber threat analysis should consider:

Figure 1: Adversarial pathway to an attack as aid for cyber intelligence analysts (INSA 2013)

CTI is thus not a "collection discipline" but more of an "analytical discipline" that informs "decision makers on issues pertaining to all levels in the cyber domain", namely the strategic, operational and tactical (Mattern et al 2014). On a strategic level, CTI should identify the intent, capability and opportunity that actual and potential malicious actors could have (Lee 2014*a*). On a tactical level, CTI identifies network threats and informs responses. Bridging the mostly non-technical strategic and narrow technical/tactical layers, the operational level is focussed on an organisation's immediate operating environment (INSA, 2014*a*).

Moving from the conceptualisation of CTI in the preceding paragraphs, the notion of CCI and its relation with CTI are now examined.

2.2 Delineating cyber counterintelligence and its relation with cyber threat intelligence

What then is CCI, how does it differ from CTI and what is the relation between these fields? As will be shown in this subsection, CCI's focus is paradoxically narrower and broader than that of CTI. CCI is narrower in that its external dimension is directed against a very specific category of "cyber hazards", namely that of hostile intelligence actions playing out in the cyber sphere. However, CCI is also broader than CTI in several respects. CCI is for one not limited to the producing and disseminating of intelligence. It also engages internal and external threats through a wide array of offensive and defensive measures. These measures are executed in synergy in accordance with the principles of traditional, multi-disciplinary CI.

2.2.1 Demarcating counterintelligence

Therefore, CCI and its relation with CTI, can only be understood and definitively defined within the context of CI generally. CI has been discussed in some detail in earlier contributions (Duvenage & von Solms 2015; Duvenage, von Solms & Corregedor 2015). Since familiarity with the concept CI is essential for further unpacking of CCI and for contextualising the CCI framework, a brief recapitulation is provided here.

As suggested by its composite terms 'counter' and 'intelligence', counterintelligence is essentially about the countering of hostile intelligence actions. Of these hostile intelligence actions, espionage (i.e. secret intelligence gathering) is perhaps the best known example. In addition to espionage, hostile intelligence activities also can include covert action (e.g. non-attributable influencing and deception). These hostile intelligence actions target valuable bodies of information as well as the people, processes, technologies and repositories wherein it resides.

Hostile intelligence actors typically execute their actions through a combination of human ('spies') and technical means. The exploitation of the cyber sphere to realise intelligence ends is part of such technical means.

The CI mission is to safeguard, but also to advance organisational strategy and assets actively. In order to execute its mission, CI has three main thrusts namely an offensive focus, a defensive focus and an intelligence function. These three dimensions constitute the CI trident. In execution of these three dimensions, CI relies on an extensive array of means, measures and methods. In traditional CI, this ranges from defensive information security measures to the offensive running of a mole or double agent. These thrusts and their relation with means, measures and methods are explained in more detail as part of the discussion on CCI.

To summarise CI can be defined as the activities conducted to "identify, deter, exploit, degrade, neutralise and protect against adversarial intelligence activities deemed as detrimental or potentially detrimental to the own interests" (Duvenage, von Solms, Corregedor 2015). Effective CI takes on, and guard against, hostile intelligence on a human (HUMINT) and technical (TECHINT) level. This technical level includes the cyber sphere as one of its conduits.

2.2.2 Defining cyber counterintelligence and its relation with cyber threat intelligence

Building on the preceding outline, CCI can be "described as that subset of multi-disciplinary CI aimed at detecting, deterring, preventing, degrading, exploiting and neutralisation of adversarial attempts to collect, alter or in any other way breach the C-I-A [confidentiality, integrity and availability] of valued information assets through cyber means" (Duvenage & von Solms 2015; Duvenage, von Solms, Corregedor 2015). As is clear from this definition, CCI shares CI's defensive and offensive missions (Bardin 2011). Defensive CCI seeks to deny an opponent the access it seeks, to guard the organisation against insider threats and vulnerability (Bodmer et al 2012). Offensive CCI's signature role is engaging and exploiting adversarial cyber actions to own advantage. It aims to neutralise a competitor's intelligence efforts through measures ranging from deception and manipulation to the degrading of adversarial cyber intelligence activities and systems (Farchi 2012, Lee 2014*b*). This exploitation can take the form of deception, disinformation and degrading. The ultimate aim of offensive CCI should be the control and exploitation of an adversary through the manipulation of its cyber intelligence action.

Effective defensive and offensive CCI cannot be executed blindly but is guided by intelligence. Similar to CTI, analysis is necessary to generate intelligence from information and data collected. Since CCI is about the outmanoeuvring of intelligence adversaries, high-quality analysis is imperative. In this regard Godson (2001) states: "Perhaps the queen of the counterintelligence chessboard is analysis – both offensive and defensive." CCI requires this high-grade intelligence on own cyber-relevant vulnerabilities (weaknesses of people, processes, facilities and technologies) actual and potential adversaries as well as on a strategic level, the macro-environment.

CCI executes it offensive-offensive missions and the collection of data and information through a wide array of measures (Bardin 2011). It must be emphasised that care should be taken not to categorise a CCI measure or methods rigidly as defensive or offensive. In numerous instances a measure can be of service to both the defensive and offensive missions. In addition, several of the offensive-defensive measures collect data or information of relevance to the CCI intelligence mission. These measures and the multi-purposes they serve are shown in the taxonomy provided in Table 1 (next page).

Table 1: A taxonomy of CCI means, methods and measures (updated and adapted from Duvenage & von Solms, 2015)

Defensive Mode		
Passive ⟷		***Active***
Deny	***Detect***	***Collect***
Physical Defensive	**Personnel/User Defensive**	**System Defensive**
Protects against: Unauthorised access to facilities and systems *In loco* theft of data, hardware	Consists of aspects such as: IT and user personnel **vetting,** re-vetting, confidentiality agreements and monitoring	Comprises a combination of: Hardware and software such as: Network perimeter-based security (filters, certain firewalls, **IDS** and **IPS** etc.) Malware scanners. Integrated automated systems/tools (that collect and evaluate information about devices connected to a

<table>
<tr><td>Introduction of malware through physical access to systems
Unauthorised altering or destruction of data
Physical destruction or access denial
Unauthorised reading (acoustic, visual, radiation, analogue, signals)
While not conventionally seen as a Physical Defence, supply-chain management has a physical defensive function. It is also part of System Defences as an enabler.</td><td>Personnel security measures, BYOD user parameters or exclusions
User programmes in cyber security that cover policy and procedures for the handling of security-related incidents, malfunctions and recovery.
Overlapping with system defences, the use of software decoys and traps to mitigate the insider threat
Investigations focussed on cyber security incidents involving personnel. Could also include digital forensic investigations.</td><td>network, activities thereon – inclusive of intrusions). Examples of such tools, discussed further on in the table, are decoys, honeypots and behavioural analyses toolsets.
Overlapping with the latter, depending on its configuration, a honeynet can be defensive or offensive in type/mode. The term fish bowling denotes the defensive configuration.
Processes (such as supply-chain management are also in part system defences).
Vulnerability assessments, penetration testing and verification testing (on products, systems, software and secure code).
Incident and threat monitoring, identification, investigation and response. A CERT is per definition defensive – although it might contain offensive elements in its responsive action.
Port level security and BYOD regulation in as far as network interfacing is concerned (Also part of Personnel Defences).</td></tr>
<tr><td></td><td colspan="2">Commercial Cyber Threat Intelligence products, services and platforms.</td></tr>
<tr><td rowspan="2"></td><td colspan="2">The use of software decoys to mitigate the insider threat is an overlap between personnel and system defensive measures. They are mostly active CCI means.</td></tr>
<tr><td>Investigations focussed on internal cyber security incidents involving personnel. May include digital forensic investigations.</td><td>Investigations of external cyber intrusions could be part passive and part active system defence.</td></tr>
<tr><td colspan="3">Offensive Mode</td></tr>
<tr><td colspan="3">Passive ← → Active</td></tr>
<tr><td>Collect</td><td>Disrupt</td><td>Exploit</td><td>Destroy</td></tr>
<tr><td colspan="2">Collection of information on and the monitoring/surveillance of the cyber sphere to detect cyber adversaries and their exploitation of the cyber sphere in a manner that is not own-network restricted – (i.e. requires more than deployment of systems described under defensive mode). Could, depending on configuration also include IDS/IPS, honey-client applications (as opposed to host-based honeypots), luring and some forms of data mining.
The recruitment and handling of virtual agents on underground forums (under true or false flag) that can serve the purpose of enticement, collection and/or exploitation. (Under certain circumstances virtual agents can also develop into HUMINT assets).</td><td>Measures taken to exploit and neutralise adversaries activities in the cyber sphere:
System and honeynet configured offensively with the aim of enticing, exploiting and deceiving adversaries. False information is displayed to adversarial reconnaissance tools, network scanners and listeners, etc. This has as one of its aims to lead adversaries in the direction of your own preference.
Utilisation of virtual agents for offensive purposes.</td><td rowspan="2">Cyber warfare, in the full extent of the term, is typically excluded from the mandate of civilian intelligence communities. A cyber warfare capability should be flexible and allow utilisation without, or in conjunction with, kinetic war.

Nevertheless, a top class civilian CCI outfit will need to have the authority and capacity to very selectively conduct operations that have cyber warfare characteristics, utilising cyberwarfare- related techniques. Such cyber CCI operations will share characteristics with covert action. (Covert action aims to influence role-players, conditions and events without revealing the sponsors identity.)

Within business, the use of offensive measures will be determined by the legislative and regulatory framework within which the entity operates.</td></tr>
<tr><td colspan="2">Cyberespionage on adversaries. Distinguishable from own-system collection (IPS, IDS, honeynets etc) on the basis that adversarial networks are targeted actively and exploited in accordance with strategic and operational objectives.</td></tr>
</table>

The preceding discussion and table show that CCI, in contrast to CTI, is not only about the delivery of intelligence products. It includes active and passive measures instituted as part of an integrated approach. Moreover, the intelligence that CCI generates covers a scope significantly wider than the actor-centric intelligence associated with CTI. From both these perspectives, CTI can thus be posited as a constituent part of CCI (cf Lee 2014*b*). Figure 2 – that should be read with the qualification on the term 'intelligence in subsection 2.2 – depicts this relationship graphically.

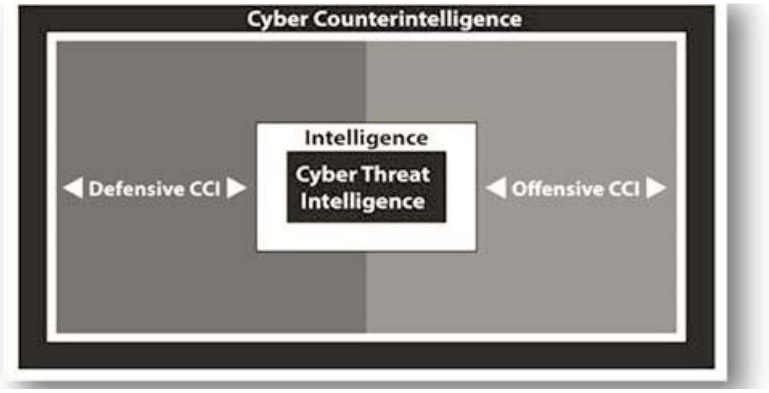


Figure 2: The relationship between cyber counterintelligence and cyber threat intelligence (authors)

This section showed CCI as a multi-faceted CI sub-discipline that participates in, but extends beyond conventional cyber security. CCI was concluded to include CTI but to be much wider in respect of scope and nature of measures undertaken.

3. Towards a multi-layered CCI framework

Effective CCI is not only multi-facetted, but also stratified. To be optimal CCI needs to involve all organisational layers from the C-suite to line-functionaries. The levels conventionally ascribed to statutory intelligence –namely strategic, operational and tactical – provide a useful approach for explaining CCI. Although these levels are described in literature dealing with CTI, no postulation could be found in open-source literature on a multi-layered framework for CCI. Works of note in the CTI field include those by Mattern (et al 2014), Friedman & Bouchard (2015), Chismon & Ruks (2015) as well as a series of papers compiled by the Intelligence and National Security Alliance (INSA 2011, 2013, 2014a, 2014b, 2015). The cited works were foundational to the framework provided in Table 2 and were also applied for the subsequent narrative description of the CCI levels.

Table 2: A multi-layered CCI framework

	Strategic	Operational	Tactical/Technical
CI mission	Advance and protect organisational interests through defence against and the offensive engagement of adversarial intelligence activities. This is achieved through the following functions: detect, deny, deter, deceive, degrade, and/or disrupt.		
CCI mission	As above, when the adversary uses cyber as a conduit or a cyber asset is a target.		
Leadership	C-level	Senior & Middle Management	Line and team leaders
Interface with CI	Organisational, Intelligence and CI Strategies All-source CI feed	Multi-disciplinary programmes and operations	Multi-disciplinary projects and continuous line-functional interaction
Referent objects	Organisation's 'crown jewels' Critical information and cyber-assets sought (e.g. adversary's 'crown jewels') Conditions (competitive advantage)	People, processes, systems, procedures (personal security, ICT architecture, supply-chain management) Own intelligence programmes	Systems, networks, and devices Network Security operations operation C-I-A (confidentiality, integrity and availability)
Interrogatives	Who, why?	Who, Where, When, How?	What, How?
Adversarial progression (Impact chain)	Motivation, intent and decision, objective	Objective Avenue of Approach Capability or perceived capability, develop access	Develop network access, implement, assess, restrike Payloads and payload delivery mechanisms
Level of adversarial role-player focussed	Sponsors, opponents, Intelligence capacity	Intelligence structures, groups, campaigns	Individuals, TTPs, incidents, actions (on-the-network)

	Strategic	Operational	Tactical/Technical
Indicators of targeting and compromise	Geo-political, sector/industry 'flags' Analogous events Adversarial strategy and business decisions	Operational disruption Organisational and/or revenue decline Information leakage	Breach in the CIA of cyber and / or information security milieu Identification of malicious code, intrusion, threat exploitation
Analysis output	High-level, strategic appraisals Strategic warning and advisories	Operational reports (CCI operations, threat, damage and vulnerability assessments, alerts, warnings) Trend analyses	Tactical and technical information reports Alerts and warnings
Consumers of CCI products	C-Level and operational management (selectively)	Line-functional managers, CI analysts and CCI specialists.	CCI analysts CCI technical personnel
Means, methods and measures (Offensive, defensive and collection)	Multi-discipline CI Strategic direction of means, methods and measures in Table 1.	As in Table 1 Interlocked with operational and tactical CI.	
Cyber threat intelligence (Sourced)	White papers, commissioned and non-commissioned research.	Platforms.	Data feeds.
Skillsets required (Line-functional)	Sound knowledge of business and industry Specialised knowledge and skills in Intelligence, multi-disciplinary CI and CCI Strategic analysis and management	Multi-disciplinary CI CCI operational and/or technical specialisation Operational management Elements of both strategic and tactical	ICT, information security Systems, software develop-ment, programming, scripting, Ethical hacking. CI and CCI tactical /technical specialisation (also HUMINT) Technical cyber defence and collection Social sciences, languages Engineering and Reverse Engineering

Within the confines of a conference paper, the framework above cannot be discussed in detail. Not even each of the vectors can be narratively explicated. The subsequent sub-sections thus do not rigidly mirror the table, but rather aim to provide a bird's eye view of the different levels on which CCI is executed.

3.1 Cyber counterintelligence on the strategic level

In his benchmark work, Prunckun 2012 rightly asserts "executive responsibility" as CI's "first and highest tenant". For CCI to be successful, the organisation's executive management (C-suite) need to understand and sanction CCI's mission to advance and protect organisational interests through defence against and the exploitation of adversarial, cyber-related intelligence activities (cf INSA 2014*b*, Chismon & Ruks 2015). Practically, the C-level executive assigned with leading the CI aspect will be responsible for also directing the CCI effort. The executive's responsibilities include obtaining the collective executive management's approval of CCI strategy, priorities and resourcing. In some instances the executive would selectively also seek endorsement – normally from the CEO – for high-risk and high-cost programmes. The actual CCI work on a strategic level is performed by a team consisting of seasoned CCI specialists, multi-disciplinary CI specialists, strategic analysts (business and CI) and various other experts relevant to the organisation's core business.

CCI informs the C-suite mainly through high-level products and presentations that include estimates, threat and risk assessments as well as advisories. These products are informed by appraisal all-source CI operational reports as well as an extensive all-source scanning of the macro-environment for CCI-relevant trends and drivers that could affect the organisation (INSA 2014*b*). External CTI products sourced would mainly be white papers as well as commissioned and non-commissioned research papers (Chismon & Ruks 2015). A thorough knowledge of organisational strategy and planning is imperative, as is a clear grasp of the organisation's information-related assets critical for it to exist and prosper – commonly referred to as the 'crown jewels' (INSA 2014*b*). It is these assets that CCI protects from adversarial intelligence activities and it is the organisational strategy that CCI should advance through the exploitation of adversaries in the cyber sphere.

Strategic CCI differs from that in the operational and tactical level in that it takes a wider view of the macro-environment and a longer term view on the actual or potential emergence of threats (Bodmer et al 2012, Mattern et al 2014). Strategic CCI would for instance identify intelligence principals/sponsors who have plausible motive, intent and capacity to target the own organisation through cyber means. (See Table 2 – "Adversarial Progression") These principals or sponsors will not necessarily execute the actual intelligence activities but are they are the ultimate benefactors (such as a nation state). The actual implementers of hostile cyber as well as associated tactics are those that carry out the task of operational and tactical CCI. While the implementers will determine the operational and tactical avenue of approach, the strategic decisions (e.g. to pursue objectives via human and/or technical means) in this regard will be taken by the Intelligence principal. The pathway of adversarial progression guiding CCI therefore differs from that of CTI (compare Figure 1).

Strategic CCI is furthermore tasked with detecting high-level indicators that the organisation is being targeted or has been compromised. Similarly, strategic CCI should identify drivers and trends suggesting a rise in the risk of internal compromise (insider threat). Equally important is the detection that organisational strategy and decision-making are being unduly influenced by deceptive, adversarial cyber operations. Strategic CCI will advise on countermeasures to best exploit adversarial cyber activities. To be successful cyber counter-deception and exploitation have to be fully synchronised with such actions in other CI fields (such as agent and double agents operations). Therefore, it is imperative for CCI to ensure that countermeasures are aligned with CI and organisational strategy. The design and filling of honeypots on the operational and tactical levels, for example, will ultimately be informed by strategic CCI's direction on counter-deception (cf Bodmer et al 2012).

3.2 Cyber counterintelligence on the operational level

As on the strategic level, CCI on the operational level strictly pursues the CCI's central mission of defensively and offensively advancing CI-relevant interests in the cyber sphere. Adherence to the mission at all three levels CCI ensures a coherent approach and an optimised CCI effort.

Operational CCI is driven by senior and middle management as well as specialists in the field of CCI operations and analysis. It functions as conduit and advisory to C-Level leadership in matters such as CCI strategic objectives, financials, financial projections and other resource requirements, projects, statistics and reporting.

Operational leadership is responsible, among other, for the following main functions (INSA 2014*a*, Mattern et al 2014): (i) operationalise the CCI strategy as set jointly by the executive management, operational management and CCI experts, (ii) develop and implement CCI structures and acquire resources, (iii) develop and implement operational plans and identify focus areas and (iv)drive daily operations and performance.

Operational CCI is responsible for safeguarding the people, processes, procedures and systems in which the organisation's critical cyber-related assets reside. Consequently, it includes a wide spectrum of organisational functions such as personal security, physical security, procurement, supply chain management, ICT-user management and much more. In addition to conducting CCI operations against adversaries (discussed below), it safeguards the organisation's own information and cyber intelligence operations. It provides operational cyber counterintelligence reports on operations, cyber threats and threat actors, damage and vulnerabilities (as identified through assessments), alerts, warnings and trends to the strategic CCI, line-functional managers, analysts and CCI specialist (Riley 2015). It also self-analyses the reports' output with a view to driving reports' outcomes to action (INSA 2013, 2014*a*).

Operational CCI interfaces with the larger CI function through multi-disciplinary programmes and operations, specifically focussing on the cyber part of CI. Its main concern is whom the adversaries are, their location, capabilities (such as the ability to utilise or develop malware), intentions (either pronounced or unpronounced) and modus operandi (Chismon & Ruks 2015). Together with this, it is concerned with the adversaries' intelligence structures and their intelligence campaigns (either planned or existing).

With regard to a traditional defensive approach, CCI similarly has a dual proactive-reactive focus to identify indicators of cyber targeting and compromise. Such indicators include a disruption in the organisational operations, tell-tale declines in organisational functionalities and/or information leakage. From a reactive perspective, CCI seeks to counter such instances by identifying its origin and addressing the compromise

(through either defensive or offensive means). From a proactive approach, it identifies such possible capabilities and campaigns and addresses threats (by either defensive or offensive means)(Bardin 2011).

Operational CCI is persistently seeking exploitable opportunities presented by adversarial cyber campaigns, operations and actions. Through counter-operations these opportunities are pursued either pro-actively or re-actively – depending on the circumstances.

The skillsets required to capacitate operational CCI are multi-disciplinary and include elements such as general management, advanced operational management, CCI analysis, cyber security, cyber defence and offensive CCI techniques and other fields of technical expertise (Bodmer et al 2012).

3.3 Cyber counterintelligence on the tactical and technical levels

The aim of tactical and technical CCI is to achieve the organisations CCI mission though tactical and technical means. It is driven and executed by line-functional leadership as well as team leaders, role leaders, CCI technical and tactical experts, security analysts and other technical personnel. It has an advisory responsibility to both the operational and executive management that includes matters such as CCI threats and opportunities, defensive and offensive measures, systems and toolsets, CCI analyses and financials (Riley 2015; INSA 2013, 2015). This advisory responsibility is usually fulfilled through submitting tactical products to the operational and in some instances directly to the strategic CCI level. Prior to submission to the executive, tactical CCI inputs are normally contextualised at the operational and strategic levels.

Tactical CCI is responsible, among others, for the following main functions (cf INSA 2015): (i) concretise operational direction into action; (ii) identify, design and implement systems, toolsets and reporting mechanisms (both defensive and offensive), (iii) carry out tactical taskings through combined technical and HUMINT measures and (iv) identify, analyse and action CCI threats and opportunities.

Tactical CCI performs the daily management, configuration (including identification and/or compromise in the case of offensive measure implementation) of both defensive and offensive systems, networks, devices, network operations and security operations (INSA 2015). It is responsible for ensuring the C-I-A of the organisation's cyber and information security environment, as a defensive tactic and measure. In the case of an offensive or exploit tactic (that must be congruent with operational objectives and the organisational strategy) tactical CCI further strives to degrade the C-I-A of an adversary's cyber and information security. Tactical CCI interfaces with the larger CI function through multi-disciplinary projects and continuous line-functional interaction. Tactical and operational CCI has a shared focus on on-the-network threats and/or opportunities, threat actors' capabilities or possible capabilities as well as the deployment of and expansion of capabilities. Tactical CCI is concerned with engaging individual groups or individuals, their specific network actions, TTPs and specific technical issues such as malware signatures (Chismon & Ruks 2015).

Tactical and technical CCI processes feed into information reports and focus on specific issues such as breaches, the identification and/or creation of malicious code, intrusion, threat and exploitation. The process leads to the compilation of tactical and technical reports, alerts, warnings, defensive and offensive solution and action reports, campaign proposals, etc. These are provided to CCI analysts, tactical leadership, operational leadership and the executive in the manner described above (Friedman & Bouchard 2015).

The skillsets required for tactical and technical CCI are, as is the case with strategic and operational CCI, multi-disciplinary. They include elements of tactical and line-functional management, ICT security, development of systems and software, programming, scripting, developing offensive and defensive toolsets, CCI technical specialisation, HUMINT and intelligence collection, as well as language and social science expertise (used in for example penetration of hacking forums), ethical hacking, technical defensive and offensive measures as well as reverse engineering (Bodmer et al 2012).

4. Conclusion

This paper emphasised the centrality of CCI to engage morphing high-end cyber threats effectively. Although only well-resourced entities can afford a fully-fledged capacity in this field, a CCI mindset and approach could benefit smaller organisations. Within the context of CCI's infancy as an academic discipline, the paper sets out to contribute two further conceptual building blocks, namely a CCI–CTI differentiation and a multi-layered

framework. Constructs such as these are important since they condition our approach to the practice. Since considerable further research is required, both constructs presented are qualified as tentative soundboards intended to stimulate future debate.

There is no consensus on definitions of CCI and CTI and this paper's differentiation is inevitably contestable. It nonetheless offers a start. The framework explicated activities on different organisational levels. As it stands, it provides more clarity on what CCI is and what it is supposed to do. With further research this framework can be developed to a scalable template for the practical execution of CCI on all organisational levels.

Acknowledgements

The research presented in this paper forms part of a project at the Centre for Cyber Security (Academy for Computer Science and Software Engineering, University of Johannesburg) aimed at formalising CCI as a multi-disciplinary field of academic inquiry in the South African context. Those interested are invited to contact the authors and/or view more detail at http://adam.uj.ac.za/csi/CyberCounterintelligence.html.

References

Bardin, J. (2011) "Ten commandments of cyber counterintelligence", *CSO* [online], http://www.csoonline.com/article/2136458/

Bodmer, S. A. et al (2012) *Reverse deception–Organized cyber threat counter- exploitation*, McGraw-Hill, New York.

Chismon, D. and Ruks, M. (2015), *Threat Intelligence: Collecting, Analysing, Evaluating*, MWR Infosecurity, UK Cert, United Kingdom.

CrowdStrike (2016) *Global Threat Report 2015* [online] www.crowdstrike.com/global-threat-report-2015/

Deloitte (2014) "Cyber threat Intelligence: Moving to an Intelligence-driven cybersecurity model." *Insight* , CIO edition, [online] http://www2.deloitte.com/content/dam/Deloitte/lu/Documents/risk/lu-cyber-threat-intelligence-cybersecurity-29102014.pdf

Duvenage, P. C. and von Solms. S.H. (2015) "Cyber Counterintelligence: Back to the Future", *Journal of Information Warfare*, Vol. 13, Nr 1.

Duvenage, P.C, von Solms, S.H. and Corregedor, M (2015) "The Cyber Counterintelligence Process - a conceptual overview and theoretical proposition", Paper read at the 14th ECCWS, Hatfield, United Kingdom, July.

Duvenage, P. C. and von Solms. S.H. (2013) "The Case for Cyber Counterintelligence", Paper read at the 5TH Workshop on ICT Uses In Warfare and the Safeguarding of Peace, Pretoria, South Africa, November.

EMC^2 (2014) *Intelligence Driven Threat Detection and Response (White paper)*, [online], https://www.emc.com/collateral/white-paper/h1304-intelligence-driven-threat-detection-response-wp.pdf

Friedman, J. and Bouchard,M. (2015) Definitive Guide to Cyber Threat Intelligence, [online], https://cryptome.org/2015/09/cti-guide.pdf

Godson, R. (2001) *Dirty tricks or trump cards - U.S. covert action and counterintelligence*. Transaction Publishers, New Brunswick.

Google (2016), Search "threat+intelligence", (2016-02-16)

Info-security Magazine (2015) "Global threat intelligence services spending is projected to rise", [online], http://www.infosecurity-magazine.com/news/cybersecurity-spending-to-hit/

INSA -Intelligence and National Security Alliance (2015*a)*, *Tactical Cyber Intelligence*, online, http://www.insa online.org/i/d/a/b/TacticalCyber.aspx

INSA(2014 *a)*, *Operational Cyber Intelligence,* [online] www.insaonline.org/i/d/a/b/OCI_whitepaper.aspx

INSA (2014 *b) Strategic Cyber Intelligence*, [online] www.insaonline.org/i/d/a/b/StrategicCyberWP.aspx

INSA (2013) *Operational Levels of Cyber Intelligence*, [online], http://issuu.com/insalliance/docs/ insa_wp_ cyberintelligence_pages_hir/16?e= 6126110/4859250

INSA (2011) *Cyber Intelligence: Setting the Landscape for an Emerging Discipline*, [online] www.oss-institute.org/storage/.../insa_cyber_intelligence_2011.pdf

iSightpartners (2014) *What is Cyber Threat Intelligence and why do I need it?* [online], http:// www.isightpartners.com/wp-content/uploads/2014/07/iSIGHT_Partners_What_Is_20-20_Clarity_Brief 1.pdf

Kaspersky (2015) *Global IT Security Risks Survey 2015: The current state of play* [online] http:// media.kaspersky.com/en/business-security/it-security-risks-survey-2015.pdf

KPMG (2013) *Cyber threat intelligence and the lessons from law enforcement*, [online] http://www.kpmg.com/Global/en/IssuesAndInsights/ArticlesPublications/Documents/cyber-threat-intelligence-final3.pdf

Lee, R. M. (2014a) "Cyber Threat Intelligence". *Tripwire* ,blog series, part 5. Retrieved on 04 January 2015 from http://www.tripwire.com/state-of-security/security-data-protection/cyber-threat-intelligence/

Lee, R. M. (2014b), "Cyber Counterintelligence: From Theory to Practice". *Tripwire*, blog series, part 4. Retrieved on 04 January 2015 from http://www.tripwire.com/state-of-security/.../cyber-counterintelligence-from-theory-to-practice/

Mattern, T. et al (2014) "Operational Levels of Cyber Intelligence"*, International Journal of Intelligence and Counterintelligence*, vol. 27, no. 4.

McAfee Labs (2015) 2016 Threats Predictions http://www.mcafee.com/us/resources/reports/....predictions-2016.pdf
Moyo, A. (2015) "Syndicates wreak havoc in cyber space", *ITWeb*, [online] http://www.itweb.co.za/index.php?option=com_content&view=article&id=143480:Syndicates-wreak-havoc-in-cyber-space&catid=234
PwC (2016) *Global Economic Crime Survey 2016: The UK*, [online], http://www.pwc.co.uk/gecs
Prunckun, H (2012) *Counterintelligence: Theory and Practice*, Rowman & Little Publishers, Plymouth.
Riley, S. (2015) Insights *to Modern Threat Intelligence*, online, https://www.linkedin.com/pulse/insights-modern-cyber-threat-intelligence-shawn-riley?articleId=7011683228767036224
Schoeman, A. (2015)"Demystifying Threat Intelligence", *Infosecurity Magazine*, [online], http://www.infosecurity-magazine.com/opinions/demystifying-threat-intelligence/
SpearTip (2015) Cyber *Hunt Team Operations and Counterintelligence*, [online] http://www.iopw.com/Article/9461/Business--Professional-Services/Cyber-Hunt-Team-Operations-and-Counterintelligence?gPage=60
Statista (2016) *Threat Intelligence Services Worldwide*, [online], www.statista.com/statistics/417588/threat-intelligence-spending/...../
The Economist (2015) "*Counter-intelligence techniques may help firms protect themselves against cyber-attacks", [online],* http://www.economist.com/news/business/21662540-counter-intelligence-techniques-may-help-firms-protectthemselves
VeriSign (2012) *Establishing a Formal Cyber Intelligence Capability,* (White Paper), [online], https://www.verisigninc.com/assets/whitepaper-idefense-cyber-intel.pdf.

A Semantic web Approach for the Organisation of Information in Security and Digital Forensics

Dagney Ellison and HS Venter
University of Pretoria, South Africa
dagneye@hotmail.com
heinventer@gmail.com

Abstract: The possession of information alone is not enough; one needs adequate access to the information. Whilst many own a book or a device connected to the internet, quick and efficient access to sought information is not guaranteed. This delay can have negative ramifications when time is critical in combating a digital incident. This paper looks at one possible means of shortening this delay by providing a solution where information discovered in the fields of digital security and digital forensics can be updated and maintained as well as easily accessed. The chosen means in addressing this problem is a knowledge base – a repository of information with the ability to grow and change along with advancements in technology. Such a knowledge base functions as a type of library with which potential solutions to specific digital forensic problems may be revealed rapidly and with immediate relevance. The semantic web is a knowledge based system; here is proposed the various layers of a semantic web constructed for the fields of digital security and digital forensics. A key feature of storing concepts in a knowledge base is that data can be directly linked to data instead of entire documents linked to documents. Furthermore, terms, concepts and assumptions used within the domains of digital security and digital forensics are made explicit and therefore facilitate better decision making in responding to a digital incident. The first step in creating a knowledge base is the development of an ontology in order to abstractly represent the chosen domains of digital security and digital forensics. The research draws on ontologies that already exist within these fields and the methodology followed during the ontology development is detailed. The implementation of this ontology is called a knowledge base and makes available a database which can be queried against in order to find the most relevant information given a set of inputs – depending on the digital investigative case at hand – and detailing the features and uses of techniques, tools and methodologies, along with other resources.

Keywords: ontology, semantic web, knowledge base, digital forensics, security

1. Introduction

There is a wealth of information available at present within the domains of information security and digital forensics. These two domains are but a subset of all the information available to us today – much of which can be found online. The internet is an expansive repository of resources, web pages and web services.

Information stored online currently exists in the form of linked documents which can be discovered through search engines. Search engines index key words and search according to the user's entered search terms (Franklin 2000). Documents retrieved from a search contain hyperlinks to other documents. This has been the most common means, up until now, by which related information has been shared and linked digitally. Linking information in this manner is not optimal as the connection is undefined and therefore does not promote any sort of sophisticated determination of the information that is sought. However, there might be a better way – that of the brainchild of Tim Berners-Lee called the semantic web (Berners-Lee et al. 2001).

The information that exists presently in the fields of information security and digital forensics is abundant and broadly dispersed amongst various academic publications, websites and other mediums. Furthermore, information that was applicable for a particular technology previously may not be relevant anymore as cyber criminals are constantly finding new ways to hide malicious code. Yet that relevance is not updated anywhere but rather hopefully inferred through generally available information. Thus, the goal undertaken in this paper is to establish the beginnings of a formalised body of knowledge for the fields of information security and digital forensics.

There are many components that make up a semantic web. The development of these components is addressed in the background section, section 2. Following this, an outline of the contribution to a formalised body of knowledge is given in section 3. Section 4 details a critical evaluation of the given contribution. Section 5 considers work related to the development of a semantic web for a particular domain of knowledge and finally section 6 concludes this paper.

2. Background

A semantic web allows for additional items within information security and digital forensics to be easily integrated and interoperated. A semantic web functions by linking data within documents to data within other documents based on a predefined set of vocabulary and semantics (Crowther 2008). This is depicted in Figure 1 below.

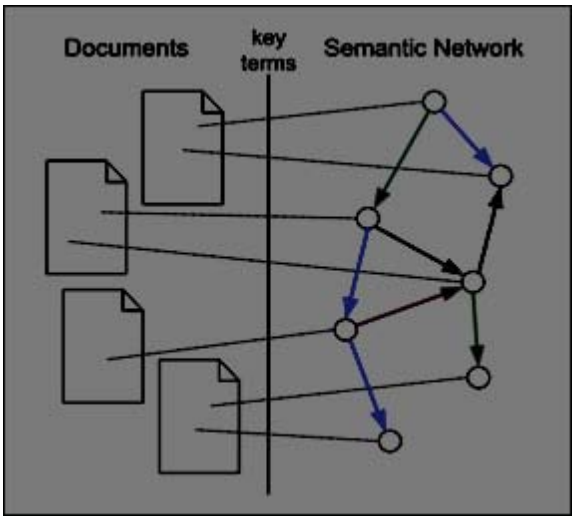


Figure 1: Semantic network of terms situated in the Web that become a semantic web (González 2005)

The semantic web is an extension of the current web and operates by creating meaningful links between information – not just documents. Relevant data can more readily and easily be discoverable as information in a semantic web is given well-defined meanings. In this way computers and people are better enabled to work in cooperation (Berners-Lee et al. 2001). The semantic web is made up of a stack of standards and implementation layers integrated to make referencing specific data achievable.

2.1 The semantic web architecture

Semantic web development comprises a variety of steps, tools, languages and standards. In a nutshell, "Ontologies are part of the W3C standards stack for the Semantic Web, in which they are used to specify standard conceptual vocabularies in which to exchange data among systems, provide services for answering queries, publish reusable knowledge bases, and offer services to facilitate interoperability across multiple, heterogeneous systems and databases. " (Gruber 2009). The main focus in this paper is the ontology in order to define the concepts and relationships between the concepts for the domain at hand. The domains information security and digital forensics are merged in this ontology as digital forensics can be considered a sub-set of information security.

The ontology is the conceptual layer of a semantic web. An ontology is an abstract entity that represents the domain of knowledge. An ontology reveals nothing about the manner in which information is stored physically but abstractly defines the domain. An implementation of the ontology is a physical entity and a type of database or repository called a knowledge base. "Ontology engineering is concerned with making representational choices that capture the relevant distinctions of a domain at the highest level of abstraction while still being as clear as possible about the meanings of terms." (Gruber 2009) – this is a quote by Tom Gruber and is what we are doing here.

The following figure, figure 2, depicts the basic layers of a semantic web.

Starting at the bottom and moving upwards, URI provides global identifiers, in other words links to data resources, and Unicode supplies a character-encoding set that allows for international characters. Above these is the Extensible Markup Language (XML) layers. XML at the bottom provides a common syntax for the semantic web being implemented. It contains an XML namespace – a library – to allow for broader vocabularies to be used within the domain. (Rhizomik n.d.)

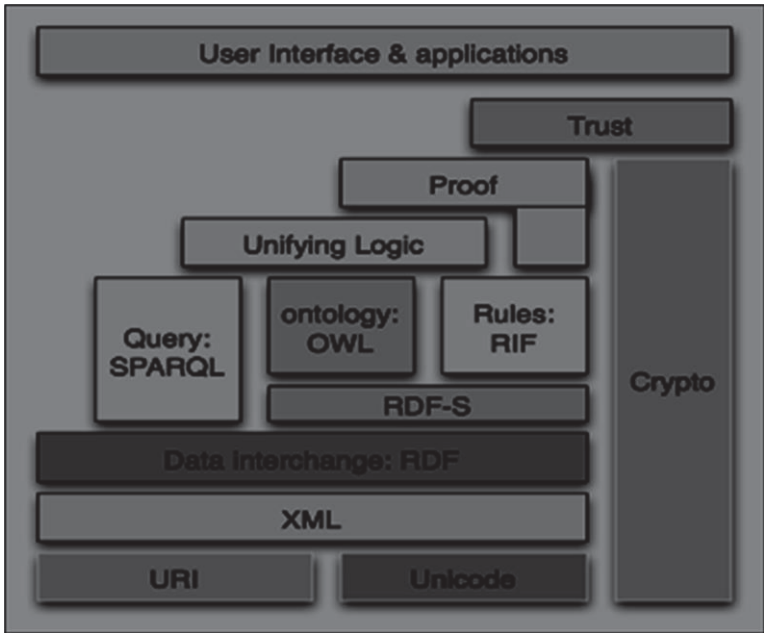


Figure 2: Semantic web basic architecture (Feigenbaum 2006)

Above XML is the Resource Description Framework (RDF) used to represent the resource items stored in the ontology in the layer above. RDF stores the triples of the ontology. A triple is a pair of nodes with a relationship between them. This is in the form of subject-predicate-object. Each node (subject or object) is a class and each relationship (predicate) is a property. On top of RDF is the RDF Schema (RDFS) which contains formal semantics in order to describe the classes and properties that exist within the ontology. (Obitko 2007)

Web Ontology Language (OWL) is the namespace used specifically for ontologies developed in a semantic web type of context. OWL provides even more vocabulary options and is embedded into RDF. It is against OWL that an ontology query language may be executed in order to deduce the most appropriate result for an agent - human or computer – to obtain relevant information. This querying can be done with a query language such as Simple Protocol and RDF Query Language (SPARQL). (W3C 2013)

Alongside ontology and query is rules. Rules place restrictions and constraints on the ontology which is useful in querying and reasoning about the data stored within the ontology. It is specifically designed to query data that is stored in RDF format (W3C 2008).

The layers mentioned are those specifically relevant for understanding in this paper. The remaining layers are applicable in a final semantic web implementation.

This semantic web structures is what allows for data to be added or modified easily. This is important for this research as technology will continue to grow and therefore the changes and additions that will be needed to be made to these fields must be maintained.

2.2 Semantic web ontology development process

The development methodology chosen for this semantic web implementation is called Methontology – a blend of the terms methodology and ontology as claimed by (Fernández-López et al. 1997). This methodology was chosen as it provides a well-structured methodology to build ontologies from scratch. The main focus of this methodology is to build the ontology in the right sequence of steps. In summary, the methodology is as follows:

- 1. Specification – at least a semi-formal specification written in natural language. Must be concise, partial and consistent.
- 2. Knowledge Acquisition – E.g. through interviews or text analysis.
- 3. Conceptualisation – structuring the domain of knowledge.
- 4. Integration – consider reusing other already existing ontologies.

- 5. Implementation – coded in a formal language.eg Java, C++.
- 6. Evaluation – a technical judgement of the ontology.
- 7. Documentation – for the sake of knowledge reuse.

In addition to these points, the process of developing an ontology is an iterative one. The purpose for which the ontology is being created must be kept in mind and there is more than one correct solution to model a domain (Noy & McGuinness 2001)

The Methontology development process described above has been applied in our preliminary ontology revealed in the next section. Due to space constraints the finer details of the Methontology have been omitted.

3. InfoRensic ontology

This section presents the starting point in the creation of a body of knowledge for digital forensics and information security by means of an initial, top-level ontology.

The ontology is given in figure 3 below. The ontology is built with the aim of expanding. As this is just a top level model, it is designed with the idea in mind that one of these nodes could be taken and an entire sub-ontology created for that sub-domain and merged back into this ontology for an even more comprehensive solution. We call this ontology the InfoRensic ontology.

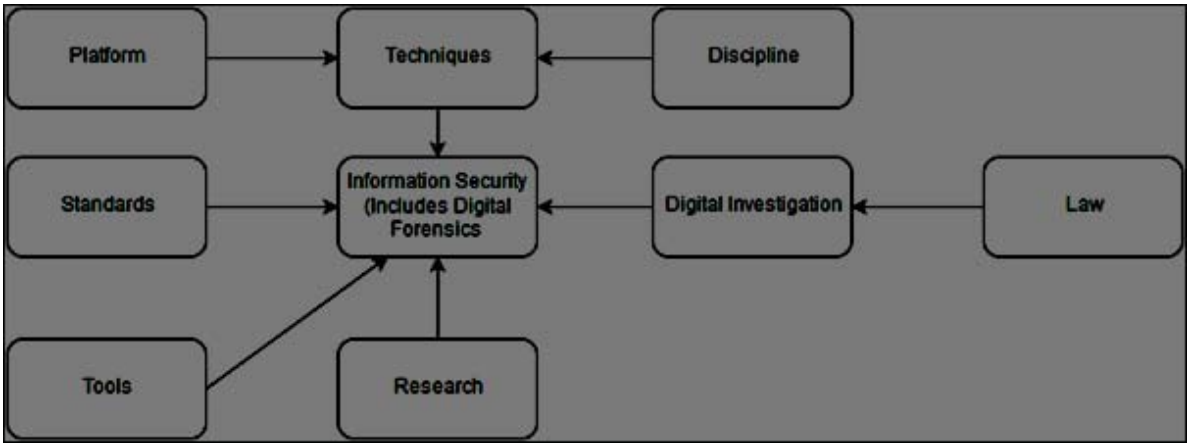


Figure 3: Proposed top-level inforensic ontology for a semantic web on digital forensics and information security

The InfoRensic ontology presented in figure 3 contains nine nodes (entities). It aims to contain all the components within Information Security and Digital Forensics in such a way that new information within the domain can be added with ease. This is the top-most level of the ontology. Further work on specific entities such as *Techniques* (Ellison & Venter 2016) has already been accomplished and could be merged here. The addition of an ontology on techniques would allow for deeper reasoning of the ontology when queried against. This would help to locate more specific information on the techniques and their application.

The entity *Information Security* has in brackets includes Digital Forensics as digital forensics is a subset of information security. The term *Information Security* was chosen as a more descriptive means of depicting security issues pertaining to the protection and preservation of digital information than just security. There are five entities which pertain to *Information Security*, namely *Tools*, *Standards*, *Techniques*, *Research* and *Digital Investigation*.

Tools represents the currently existing tools available for both preparing against a digital incident as well as responding to a digital incident. This could include, for example, Forensic Tool Kit (FTK). The *Tools* entity could be classified further with subsumption (inheritance) relationships to enable more specific query results. An example of further classification would be to have two entities under *Tools* called *Proactive Tool* (a tool used to set up a preventative measure towards an incident) and *Reactive Tool* (a tool used to react to an incident, e.g. FTK). This further classification would separate the two types of tools and they would then inherit only the properties of that super-class which would be more refined.

Standards represents the published standards that are available in information security and digital forensics, for example the body of ISO Standards contains many standards pertaining to information security such as the ISO 27043 Standard. This *Standards* entity could be classified further with subsumption relationships to facilitate more specific query results. For example, the ISO 27043 Standard which acts as an umbrella standard could be classified as a type of umbrella standard. The standards referenced within the ISO 27043 Standard could be classified as non-umbrella type standards with properties indicating the standard they are referenced in. A query of the standards could therefore produce results not only of the ISO 27043 Standard but also of reference to the ISO 27037 Standard which the ISO 27043 Standard mentions.

Techniques represents the techniques that are available for preparing against a digital incident, e.g. setting up a firewall, and reacting to a digital incident, e.g. tightening firewall rules. Techniques apply on a set of platforms. *Platforms*, therefore, represents the basic details of the device upon which a technique is being carried out. It is inferred, therefore, that this is the platform upon which information security or digital forensics is being conducted in general. If no technique is occurring, nothing towards information security or digital forensics is being achieved and thus no platform is applicable. There is therefore a constraint that a technique must be performed on a platform if it should be of a technical, implementable nature. More on this is provided in the table of example metadata below.

In addition to *Platform*, *Discipline* also pertains to a technique. Information on the discipline being practiced is only applicable when a technique is being carried out. Like *Platform*, *Discipline* only applies if the technique being carried out is of a technical, implementable nature. More information is given in the table of example metadata below.

Digital Investigation is the fourth entity linked to Information Security. This is the point where information security and digital forensics has its bearing – if there were no incident, there would be no need for security or forensics. This is also the point to which *Law* applies. The law contains Acts which contain sections and ultimately clauses that become specifically applicable in a particular case where digital forensics is involved. Applicable laws can become specified in this entity of the ontology.

Research, finally, represents academic publications useful in learning more about a particular technique or tool, for example. This may include papers, articles, books etc.

The entities of the InfoRensic ontology were specifically chosen to represent all the facets of information security and digital forensics. With each facet represented, basic metadata which could be applied for any potential instance of that facet could be established. This metadata is ultimately what queries against the ontology will query against. The preliminary metadata, therefore, has been provided in figure 4 below.

Table 1: Table of initial metadata and examples based on would-be instances

Entity	Metadata labels	Metadata type	Example
Tools	Name	String	EnCase
	Version	String	7.10
	Operating System	String	Windows
	Licence	String	Proprietary
Standards	Name	String	ISO 27043
	Organisation	String	ISO
	Year	Integer	2015
	Subject	String	Information Security
Techniques	Name	String	Disk Dump
	Nature	Enum [Hard (technical), Soft (non-technical)	Hard
	Medium	Enum [HardDrive, TextFile, ImageFile, VideoFile, DatabaseFile]	HardDrive
Platform	Operating system name	String	Windows
	Operating system bit size	Integer	64
	Release	String	8.2

Entity	Metadata labels	Metadata type	Example
	Processor	String	Intel Core i5-5200U2.2GHz
	RAM	Integer	4096
Discipline	Name	String	Cloud
	Cross-platform	Boolean	True
	Cross-device	Boolean	True
Digital Investigation	Name	String	AshleyMadison Hacking Case
	Year	Integer	2015
	Country	Enum [<all countries>]	USA
	Type	Enum [Theft, Hack, Exploitation]	Hack
Law	Name	String	Protection of Personal Information Act
	Year	Integer	2014
	Jurisdiction	Enum [<all jurisdictions>]	South Africa
Research	Title	String	The Semantic Web Revisited
	Author	Array of Strings	Tim Berners-Lee, Wendy Hall
	Year	Integer	2006

In Figure 4 the entities of the InfoRensic Ontology are given hypothetical metadata based on potential instances inferred from the name attribute in the table. The ontology becomes a knowledge base once it is implemented and instances are added to the entities.

With the kind of metadata that is presented in the table in figure 4, the ontology can be queried against in order to find, for example, the details of a discipline that pertain to a specific technique. If the technique were disk dump on ROM memory, the ontology should be able to reveal that this is a dead forensics technique that is conducted on a single device and on a single platform. Once the instance is pulled up, details of the technique, such as the terminal command, can be made available. Related instances of the disk dump technique will also be listed.

All the entities and attributes in this InfoRensic ontology can be represented with the aforementioned tools RDF and OWL and queried against with the SPARQL query language. The next section presents an overview of related work of similar semantic web approaches as well as other ontologies in the field.

4. Related work

Fonou-Dombeu and Huisman undertook a similar task in creating a semantic web for an E-government domain (Dombeu & Huisman 2011). Fonou-Dombeu and Huisman focused on the ontology development aspect as well as it is the link between the abstract representation and the physical artifact of the knowledge base. Fonou-Dombeu and Huisman took a different methodology approach and used the methodology presented by Uschold and King for a semi-formal representation of the domain of e-government. Both the paper by Fonou-Dombeu and this paper aim to build a semantic web by making use of the common semantic web tools such as RDF and OWL along with ontologies. The only difference is the domain being modelled.

Brinson, Robinson and Rogers proposed an ontology that also covered information security and digital forensics at a high level (Brinson et al. 2006), however, the major difference compared with this paper is the purpose for which they created it. In this paper the aim of the InfoRensic ontology is to lay the foundation for a structured, query-able body of knowledge. Brinson, Robinson and Rogers created their ontology for specialisation, certification and education. Therefore, possible queries executed against their domain would return results pertaining to certification, specialisation and education whereas this InfoRensic ontology would return information regarding a broad range of topics within the information security and digital forensics domain, such as related techniques, relevant standards and research.

Finally, an ontology that could potentially be merged into the InfoRensic ontology is one by Karie and Venter towards a general ontology for forensic disciplines (Karie & Venter 2014). In their paper, Karie and Venter present a detailed ontology on forensics disciplines including network, multimedia and database forensics classified

into sub-disciplines down to objects and sub-objects. The aim was for better organisation in the field of digital forensics and to enhance the sharing and reuse of the formally represented knowledge in digital forensics.

The selected publications in this section relate most strongly to the InfoRensic ontology and could possibly even be merged with the ontology. The next section concludes the work carried out in this paper.

5. Conclusion

The main building block of the semantic web is the design and development of the ontology as the ontology acts as the interface of abstraction between the stored data artifacts and what the user sees and is searching for. This work presents the preliminary top-level ontology to be used in the semantic web currently being developed for the domains of information security and digital forensics. The problem of dispersed data in these fields could then begin to be overcome by the consolidation of the various aspects of these fields into this one, single semantic web of knowledge for information security and digital forensics. This work allows for information stored within the final, implemented repository to be shared and reused and also to be queried against for better relevant data retrieval. Furthermore, all the advantages of implementing an ontology become realised which includes making domain assumptions explicit as well as defining terminology for the domain.

The ontology aims to address all aspects concerned with a digital incident and pertaining to a digital investigation. The ontology should have the potential to hold any new piece of information with ease due to the easily-expandable nature of an ontology. Further work is being done on other related ontologies to develop a more fully comprehensive and inclusive system for the domains of digital forensics and information security and especially including the contributions of others where an entity in the ontology presented here may be represented more fully with a sub-ontology merged into this one.

References

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American*, 284(5), pp.34–43.

Brinson, A., Robinson, A. & Rogers, M., 2006. A cyber forensics ontology: Creating a new approach to studying cyber forensics. *Digital Investigation*, 3(SUPPL.), pp.37–43.

Crowther, R., 2008. Planning a Semantic Web site. Available at: http://www.ibm.com/developerworks/library/x-plansemantic/.

Dombeu, J.V.F. & Huisman, M., 2011. Combining Ontology Development Methodologies and Semantic Web Platforms for E-government Domain Ontology Development. *International journal of Web & Semantic Technology*, 2(2), p.14. Available at: http://arxiv.org/abs/1104.4966.

Ellison, D. & Venter, H., 2016. An Ontology for Digital Security and Digital Forensics. In T. Zlateva & V. Greiman, eds. *Proceedings of the 11th International Conference on Cyber Warfare & Security*. Boston, USA, pp. 119–127.

Feigenbaum, L., 2006. Semantic Web Technologies in the Enterprise. Available at: http://www.the-figtrees.net/lee/blog/2006/11/semantic_web_technologies_in_t.html.

Fernández-López, M., Gómez-Pérez, A. & Juristo, N., 1997. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *AAAI-97 Spring Symposium Series*, SS-97-06, pp.33–40. Available at: http://oa.upm.es/5484/.

Franklin, C., 2000. How Internet Search Engines Work. Available at: http://computer.howstuffworks.com/internet/basics/search-engine1.htm [Accessed April 5, 2016].

González, R.G., 2005. *A Semantic Web approach to Digital Rights Management by*.

Gruber, T., 2009. Encyclopedia of Database Systems. In L. LIU & M. T. ÖZSU, eds. Boston, MA: Springer US, pp. 1963–1965. Available at: http://dx.doi.org/10.1007/978-0-387-39940-9_1318.

ISO/IEC, 2013. Information technology — Security techniques — Incident investigation principles and processes. , 2014(40), pp.13–23.

Karie, N.M. & Venter, H.S., 2014. Toward a General Ontology for Digital Forensic Disciplines. *Journal of Forensic Sciences*, 59(5), pp.1231–1241. Available at: http://doi.wiley.com/10.1111/1556-4029.12511.

Noy, N. & McGuinness, D., 2001. Ontology development 101: A guide to creating your first ontology. *Development*, 32, pp.1–25. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.5085&rep=rep1&type=pdf\nhttp://liris.cnrs.fr/alain.mille/enseignements/Ecole_Centrale/What is an ontology and why we need it.htm.

Obitko, M., 2007. Translation between Ontologies and Multi-Agent Systems (Online tutorial). Available at: http://www.obitko.com/tutorials/ontologies-semantic-web/semantic-web.html.

Rhizomik, R., Knowledge Representation. Available at: http://rhizomik.net/html/~roberto/thesis/html/KnowledgeRepresentation.html [Accessed April 23, 2015].

W3C, 2013. SPARQL 1.1 Overview. Available at: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

W3C, 2008. SPARQL Query Language for RDF. Available at: https://www.w3.org/TR/rdf-sparql-query/.

An Ontology for Threat Intelligence

Courtney Falk
Optiv, Denver, USA
courtney.falk@optiv.com

Abstract: This paper describes the work done to build an ontology in support of cyber threat intelligence. The end goal is a system that helps threat intelligence analysts effectively organize and search both open source intelligence and threat indicators in order to build a comprehensive picture of the threat environment. The Lockheed Martin kill chain model serves as the basis for the ontology. Semantic Web technologies such as RDF, OWL, and SPARQL are used to leverage existing commercial off-the-shelf software and tools.

Keywords: threat intelligence, ontology, semantic web, cybersecurity

1. Introduction

Threat intelligence concerns itself with a holistic view of cyberattacks. The goals of intelligence include identifying who the attackers are, what the goals of their attacks are, and what the tactics, techniques, and procedures (TTPs) are. Actual intelligence work goes beyond collecting threat indicators by using analysis to synthesize a more complete picture.

Proper threat intelligence operates in a way similar to the work done by law enforcement and intelligence agencies (Pickens, 2015). Intelligence is a cyclical process that must begin with a set of requirements that provide scope and bounding of the problem in question. After the intelligence goals are established then collection gathers data of potential interest to the questions being asked. At this point the information is just raw data and not finished intelligence. It is in the third step, analysis and production, that trained intelligence analysts fuse data with insight into a finished intelligence product. The fourth step involves distributing the intelligence product to concerned parties and using the intelligence internally to inform changes in policy and procedures. These changes go on to inform subsequent passes through the intelligence cycle.

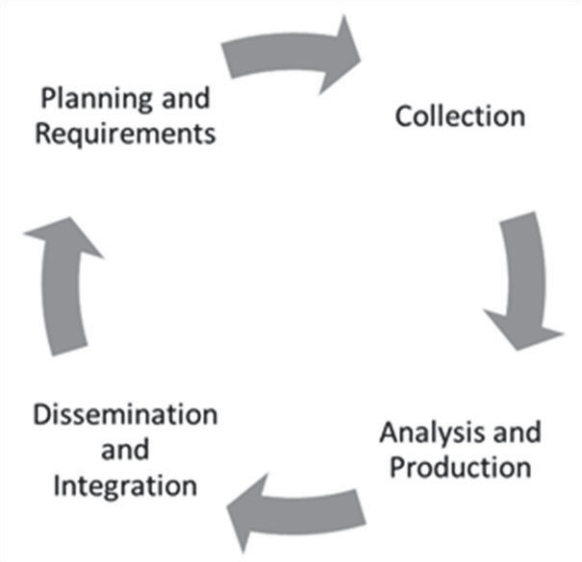


Figure 1: Demonstrating the nature of the intelligence cycle

This paper outlines the project of developing an ontology to support the work of threat intelligence. The ontology serves as a model of how the world works. In this particular case, the ontology is limited in focus and scope to problems concerning cybersecurity. The goal is to describe this threat intelligence ontology in sufficient detail for it to be of use by threat analysts in their day-to-day work.

An ontology is an abstract model of the world as it exists. It is a hierarchy of concepts with different properties connecting them together. More specific child concepts inherit these properties from their more general parent concepts. Instead of a universal ontology, this paper describes a specialized ontology focused on describing

cybersecurity concepts. This particular ontology may not be able to explain how crops are harvested or the orbital motions of planets but it doesn't need to. Instead, it is a domain ontology, dedicated to a single domain of knowledge (Guarino, 1998).

As an example of a particular domain of knowledge take the field of art. There are many media used in art; canvas for paintings, marble for sculptures, so on and so forth. Then there are the artistic techniques that utilize different media. These techniques in turn require specialized tools. An art domain ontology needs the ability to describe all these different concepts and how they interrelate. What an art domain ontology doesn't need is a description of a combine harvester, corn stalks, or grain silos. An agriculture domain ontology would be better suited for these concepts.

The Semantic Web is a framework co-designed by Sir Tim Berners-Lee that builds upon the World Wide Web (2001). Berners-Lee originally rose to prominence by designing the technology that underpinned the World Wide Web (WWW). The Semantic Web is seen as a natural evolution of the WWW. What makes the Semantic Web new and different are the semantically meaningful tags used to describe content on the web. All of these semantic tags connect back to an ontology.

An ontology intended for Semantic Web applications is implemented in the Web Ontology Language (OWL) standard (W3C, 2004). This standard is maintained by the World Wide Web Consortium (W3C) and is freely available. A popular choice for implementing OWL ontologies is the XML-based Resource Description Framework (RDF) language.

2. Model

Because ontologies are hierarchical in nature it is useful to being by describing the most abstract, high-level concepts in the ontology first. At the top of an ontology is its root, an abstract concept from which all other concepts derive. For the threat intelligence ontology there are two child concepts of the root, Object and Event. Objects are static things that exist in the world on their own. Events are occurrences that involve one or more objects. This Object/Event distinction is important when designing properties for connecting different concepts together.

Thing

- Event
- *Attack*
- *Compile*
- *Execute*
- Object
- *Persona*
- *Organization*
- *Software*
- *ComputingDevice*

The above nested list shows some of the basic divisions in the ontology. Thing is the root concept with children, Event and Object. Each of these two concepts has their own child concepts in turn. This is by no means a complete description of the final ontology.

Another distinction is created between two children of the Object concept, Persona and Organization. A persona is an individual or a personality. An individual is a specific person while a persona is an online presence projected by an individual. Because this ontology deals with online information there is no guarantee that each and every time the account or actor behind an event can be properly attributed. To that end the Persona concept is a compromise choice between the two. The Organization concept is a social group made up of personas with structure applied to the group itself. The Persona concept does not have any child concepts of its own, but the Organization concept has several child concepts to help further focus and specify new types of organizations

based on their structure or goals. The following nested list defines some basic distinctions between different kinds of organizations:
Organization

- PoliticalOrganization
- *NationState*
- *Agency*
- *LawEnforcementAgency*
- *IntelligenceAgency*
- *MunicipalArea*
- *City*
- *Town*
- BusinessOrganization
- *Corporation*
- *Partnership*
- HackingGroup

One goal of this project is to utilize existing work wherever possible. If useful models already exist then it is inefficient to design a new, similar model from scratch. To that end, the cyber kill chain model by Lockheed Martin was chosen as a starting point for the ontology. One of the benefits of using the cyber kill chain model is that it is well understood by the cybersecurity community. This familiarity helps both with the acceptance of the threat intelligence ontology project and shortens the amount of time it takes for users to learn all the necessary aspects of the model.

The cyber kill chain models cyber-attacks as a progression of seven steps (Hutchins, et al., 2011):

- 1. Reconnaissance
- 2. Weaponization
- 3. Delivery
- 4. Exploit
- 5. Install
- 6. Command and control
- 7. Actions on target

These kill chain steps are each a potential event. But before the individual steps are added to the ontology it is useful to define an Attack concept to unify them all together where Attack is a child of Event.

Next, another event, KillChainStep, is created as a sibling to Attack. KillChainStep serves to organize together all seven of the steps of the kill chain model in one place. It isn't a child of Attack because this would suggest that each of the seven steps could in and of themselves be a more specific kind attack. But this is incorrect.

Instead, the preferable solution is to define the kill chain steps as mere parts where an attack is the whole that assembles together multiple steps. The study of the relations between parts and wholes is called mereology (Green, et al., 2002). A deeper dicussion of mereology is beyond the scope of this paper.

A study of the steps of the kill chain model suggests that each step could in fact be one of several possible other events. Take the Reconnaissance step as an example. For a cyberattack, reconnaissance might consist of searching the web, querying a database, or fingerprinting a server directly. All of these events provide necessary information for continuing the attack. Weaponization, Delivery, and Exploit steps also hide underlying complexity. Creating children under these steps allows for a more fine grained description of the actions involved.

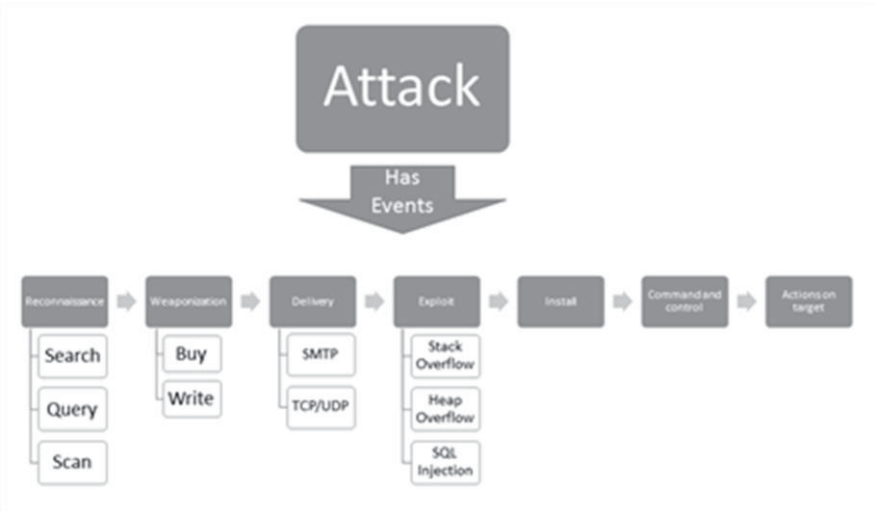


Figure 2: The diagram showing how the Attack concept relates to the kill chain concepts and on to further related concepts

OWL offers different ways to add descriptive information to the concepts in an ontology. The first was already discussed: object properties, in which both the domain and the range of the property in question are concepts in the ontology. The other type of property is a data property. With a data property the domain is a concept while the range is a non-concept value such as a scalar, Boolean, or string value.

The ontology being built here has a specific use in mind for data properties as a way to add metadata to individuals (concrete instances of the ontology's abstract concepts). Metadata properties improve searching through an ontology by adding information about the time the events happened, the source reporting the events, and perhaps even the confidence in the veracity of the information. Metadata improves reconstruction of an attack from multiple sources; conflicting information might be resolved in favor of one source over another, or different sources with different portions of the attack might be combined while retaining the knowledge of their origin.

3. Application

The paper up to this point describes an abstract ontology and some of the theory that goes into designing it. This section of the paper uses examples taken from real world situations to demonstrate some of the details of how to bridge the abstract model with concrete application.

These applications would require a human analyst to examine all the available data, and then create the necessary individuals in the ontology. This process is manpower intensive and therefore potentially expensive. Use of advanced natural language processing (NLP) software might speed this process by automatically extracting entities and relations from texts. Utilization of NLP to assist human acquisition remains a topic for future work.

3.1 Example 1: Distributed denial of service

This example uses the Phantom Squad DDoS of Xbox Live and the PlayStation Network (PSN) in December of 2015 (Walton, 2015). A distributed denial of service (DDoS) attack presents a situation that does not fully utilize the kill chain model. Steps four and later (exploit, install, command and control, and actions on target) do not occur in this example because a DDoS attack does not install or execute code on the target machines.

The first step is to create an individual called PhantomSquad that is an instance of the HackingGroup organization. Attribution of PhantomSquad as the agent of the DDoS attack is due to extensive Twitter posts made before, during, and after the attack. The content of these tweets suggested insider knowledge of the attack.

The reconnaissance stage is unconfirmed by reporting but some assumptions can be made. A simple web search could have provided the URLs for Xbox Live and the PSN because these services need to be publicly reachable.

The weaponization stage, similar to the reconnaissance stage, does not have a lot of supporting evidence. Because a DDoS attack is spread across a large number of hosts it is difficult to pick one, perform forensics on it, and attribute the bot to a specific owner. This assumption of a bot net still does not explain if PhantomSquad grew and maintained their own botnet for the purpose of conducting DDoS attacks of if they leased the services of an existing botnet.

Figure 3: The graph of individuals involved in the Phantom Squad DDoS attack

3.2 Example 2: Ukrainian power outages

The second example incorporates every step in the kill chain model. Recent attacks on the industrial control systems (ICS) of the Ukrainian energy sector succeeded in shutting off power in some areas (ESET, 2016). iSIGHT Partners calls the group behind these attacks the Sandworm Team (Hultquist, 2016). From this information, an analyst can create a new Attack individual to represent the attacks against the Ukranian ICS infrastructure as well as a new HackingGroup individual to represent the Sandworm Team. The HackingGroup is attached to the Attack using the agent object property.

The malware involved in these attacks includes BlackEnergy and KillDisk. Both of these were found on computers involved with the ICS belonging to Prykarpattyaoblenergo, which serves the Ivano-Frankivsk region. Current reports are unclear as to which piece or pieces of malware actively caused the ICS failure and subsequent blackout.

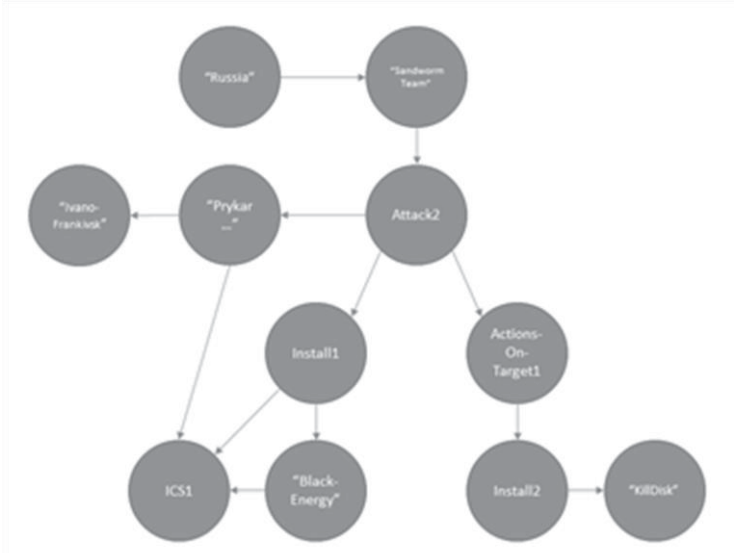


Figure 4: The graph of individuals involved in the Ukrainian blackout

Even with all this information there are still unanswered questions about the intrusion: How did the attackers identify targets, and were they performing the attacks at the direction of any other organization? The SPARQL query language is useful in this case by providing a mechanism to search through the ontology and identify gaps in the known attack sequence.

The Sandworm Team example also raises a philosophical question: how much of the physical world must an ontology devoted to cybersecurity be able to describe? Specific to this example, what are the different ways that cybersecurity failures can cause reactions in the physical world? Contrast this to the Estonia DoS attack of 2007 where daily life for citizens was affected but there was no physical damage to infrastructure.

4. Conclusion

This paper described the methods used in designing an ontology for use by threat intelligence analysts. The description of the ontology itself, as well as its components and properties, were all defined according to the formalisms used by OWL standard. OWL, RDF, and other Semantic Web components are freely available from the W3C.

The ontology for threat intelligence assists analysts by normalizing and organizing threat intelligence, and by connecting together disparate pieces of data into a single pictures. This level of organization enables the semantic search and intelligent querying of the threat intelligence.

4.1 Future work

Currently the ontology is a single, monolithic piece. A useful improvement would be to break it into three separate pieces. Separation of the parts would allow for one or more to be replaced as needed or as improvements are produced. These pieces are:

- The core ontology describing the threat intelligence model.
- Data properties useful for annotating concepts with meta-data about their source, confidence, etc.
- Threat indicators that connect to concepts within the threat intelligence model.

Each possible threat indicator standard could have its own focused ontology which imports the core threat intelligence ontology. This isolation of each threat indicator standard allows groups to use the standard most familiar to them, or to combine two or more different threat indicator standards together in one place.

The examples used in this paper demonstrate how large number of individuals are needed to be created for each attack. This can quickly become a tedious and time-consuming task for analysts and if this is too burdensome then they will not use the ontology at all. The solution to this problem is to create software tools that assist in the automated creation of new individuals. Such software would make expressing attacks in the language of the ontology as seamless as possible for the threat analysts.

References

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American,* 284(5), pp. 34-43.

ESET, 2016. *ESET Finds Connection Between Cyber Espionage and Electricity Outage in Ukraine.* [Online] Available at: http://www.eset.com/int/about/press/articles/malware/article/eset-finds-connection-between-cyber-espionage-and-electricity-outage-in-ukraine/

Green, R., Bean, C. A. & Myaeng, S. H. eds., 2002. *The Semantics of Relationships: An Interdisciplinary Perspective.* s.l.:Springer.

Guarino, N., 1998. Formal Ontology and Information Systems. *Proceedings of FOIS'98,* Volume 46, pp. 3-15.

Hultquist, J., 2016. *Sandworm Team and the Ukrainian Power Authority Attacks.* [Online] Available at: http://www.isightpartners.com/2016/01/ukraine-and-sandworm-team/

Hutchins, E. M., Cloppert, M. J. & Amin, R. M., 2011. *Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains.* s.l., s.n.

Pickens, D., 2015. *The Art of MSS Intelligence: How to establish an intelligence differentiation among competitors,* s.l.: FishNet Security.

W3C, 2004. *OWL Web Ontology Language.* [Online] Available at: https://www.w3.org/TR/owl-features/[Accessed 1 February 2016].

Walton, M., 2015. *Xbox Live pummeled by DDoS attack; hacker group claims responsibility.* [Online] Available at: http://arstechnica.com/gaming/2015/12/hacker-group-phantom-squad-takes-down-xbox-live-in-ddos-attack/

An Analytical Approach to the Recovery of Data From 3rd Party Proprietary CCTV File Systems

Richard Gomm, Nhien-An Le-Khac, Mark Scanlon and M-Tahar Kechadi
School of Computer Science, University College Dublin, Ireland
richard.gomm@gmail.com
an.lekhac@ucd.ie
mark.scanlon@ucd.ie
tahar.kechadi@ucd.ie

Abstract: According to recent predictions, the global video surveillance market is expected to reach $42.06 billion annually by 2020. The market is extremely fragmented with only around 40% of the market being accounted for by the 15 top video surveillance equipment suppliers as in an annual report issued by IMS Research. The remaining market share was split amongst the numerous other smaller companies who provide CCTV solutions, usually at lower prices than their brand name counterparts. This cost cutting generally results in a lower specification of components. Recently, an investigation was undertaken in relation to a serious criminal offence, of which significant video footage had been captured on a CCTV Digital Video Recorder (DVR). The unit was setup to save the last 31 days of footage to an internal hard drive. However, despite the referenced footage being within this timeframe, it could not be located. The DVR unit was submitted for forensic examination and data retrieval of specified video footage which, according to the proprietary video backup application, was not retrievable. In this paper, we present the process and method of the forensic retrieval of video footage from a DVR. The objective of this method is to retrieve the oldest video footage possible from a proprietary designed file storage system. We also evaluate our approach with a Ganz CCTV DVR system model C-MPDVR-16 to show that the file system of a DVR has been reversed engineering with no initial knowledge, application or documentation available.

Keywords: CCTV forensics, CCTV-DVR file systems analysis, video file carving, reverse-engineering

1. Introduction

The closed-circuit television (CCTV) is a video surveillance system that can be used for any type of monitoring. In the 2012 annual report issued by IMS Research, an independent supplier of market research, it estimated that the video surveillance equipment market was worth over $9 billion in 2010. Yet only around 40% of the market was accounted for by the 15 top video surveillance equipment suppliers (IMS Research 2012). The remaining market share was split amongst the numerous other smaller companies who provide CCTV solutions, usually at lower prices resulting from a lower specification of components. The advanced forms of CCTV normally use Digital Video Recorder (DVR) that allows CCTV video images are recorded and archived continuously from all cameras for 90 days or more, with a variety of quality. The CCTV-DVR devices are widely used for surveillance in areas that may need monitoring such as banks, airports, military installations, stores, etc.

Recently, an investigation was undertaken in relation to a serious criminal offence, of which significant video footage had been captured on a Ganz CCTV Digital Video Recorder (DVR) model C-MPDVR-16. The unit was setup to save the last 31 days of footage to an internal hard drive, however despite the referenced footage being within this timeframe it could not be located. The Ganz DVR unit was submitted for forensic examination and retrieval of specified video footage which, according to the proprietary video backup application, was not retrievable. Initial examination of the Ganz DVR unit, under forensic conditions, revealed a proprietary file system i.e. a file storage system not recognised as an industry standard. In that fashion the standard tools like EnCase (https://www.guidancesoftware.com/), X-Ways (https://www.x-ways.net/) etc. were unable to recover any video footage. In fact, DVR's come from a multitude of manufacturers, each using their own unique variants of both equipment and software. In some case this presents as an industry standard operating system, such as Linux, which presents as an easy forensic retrieval using standard tools. However in the majority of cases the forensic investigator will encounter a proprietary or drastically cobbled-together operating system which provides a significant challenge to decode. The latter is the focus of this paper.

The topic of this paper are the CCTV systems described in the previous paragraph and in particular the forensic examination of a Ganz CCTV DVR model C-MPDVR-16. The research problem was established to reverse engineer (Poole et al. 2008, Zeltser 2010) the proprietary file system and establish the details of the oldest recorded video footage available for retrieval. A secondary objective was set in being able to carve video footage from the Ganz DVR unit's internal hard disk drive into a playable format.

Our paper is set out as follows: Section 2 shows related work of CCTV-DVR forensics. We discuss on forensic techniques applied for CCTV-DVR investigations as well as forensic challenges in Section 3. We describe and discuss a case study of forensic acquisition and analysis of Ganz CCTV DVR model C-MPDVR-16 in Section 4. Finally, we conclude and discuss on future work in Section 5.

2. Related work

Ariffin et al (2013) describe a forensic technique to carve video files with timestamps. Within this paper, authors also proposed an extension to the digital forensic framework established by (McKemmish, R. 1999), which had four steps: (i) Identification, (ii) preservation, (iii) analysis and (iv) presentation. It is within Step 3: Analysis that authors specifically reference the issues with proprietary file systems. Due to the amount of variables and unknown systems available it would not be feasible to produce a technical guide to reverse engineering a proprietary file system. Instead a detailed analysis of the required steps was given; in summary the following was established: (i) First the byte storage method must be determined (little / big endian); (ii) File signatures can then be derived and used to correlate each file signature to the channel video that captured the scene with timestamps and (iii) Video coded must be located and installed. In this paper, we provides details of the analysis of the Ganz DVR unit conducted in accordance with the framework outlined by (Ariffin et al 2013). Another research was performed by Dongen (2008) on a Samsung CCTV system. This research focuses however on a well-known file system: *ext3*. In Wang (2009), author focused on Digital Video Forensics that could apply in the context of CCTV forensics.

Han et al (2015) present the forensic analysis of a CCTV-DVR of which the hard disk using a HIKVISION file system. Authors also mentioned it is an unknown file system. In this research, they identify the structure and mechanism of a HIKVISION file system. Authors show moreover that the procedure in the case analysis can be useful to counter anti-forensic activities. This paper only however focuses on HIKVISION file system that is not popular in video surveillance devices in western countries.

Recently, Tobin et al. (2014) conducted a CCTV-DVR forensics. Authors produced a short report into his examination of the file system used on the Ganz internal hard disk drive. This report was non-technical and aimed for the legal investigative function. Authors found that the Ganz internal hard disk drive was split into three regions, each of which had a specific action: (i) Region 1 – date 1 block, (ii) Region 2 – date 2 block, (iii) Region 3 – video data. To date there is any technical paper that has been published by authors on how the findings were reached; it is believed they used a software tool to watch the disk access whilst using the AvTech proprietary application, DiskTools.exe, to access a copy of the Ganz internal hard disk drive. This process has a distinct disadvantage as you are reliant on the DiskTools.exe programming, which is proven later in this paper to not provide full details of the oldest video footage available from the Ganz DVR internal hard disk drive.

3. CCTV-DVR forensics

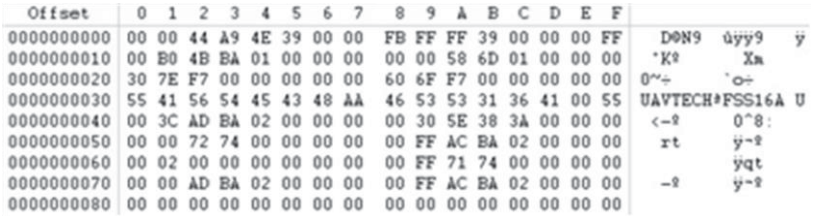
Following the literature survey in Section 2, we notice that there is very little information available on the proprietary file system used by the Ganz DVR unit, or the specified AvTech hard disk drive system. Whilst CCTV footage can and should be obtained through the proprietary application shipped with each unit, this leaves very little room for forensic examination and leaves the investigator at the mercy of proprietary programmers. Additionally there may remain remnants of video footage which are not complete and therefore not available through any such proprietary system.

By reverse engineering the AvTech file system the investigator obtains full access to all information on the DVR hard disk drive. Additionally this can be done in a forensic manner and without the need to rely on the ability of an unknown 3rd party programmer. Following the forensic acquisition a sample of video footage must be extracted from the DVR hard disk drive and converted for playback. This will require codec, headers and encoding identification.

Before looking at a case study on forensic acquisition and analysis of Ganz CCTV-DVR in the next section, we discuss on the identification of file systems of internal hard drive of Ganz CCTV-DVR (Ganz internal hard disk drive). In fact, file systems such as FAT16, FAT32, NTFS, HFS, Ext2 are some examples of industry standard file system. These file systems are used to keep track of where data is located on a disk, providing a tree like structure comprising of files and folders. Each of the industry standard file systems referenced above are required to identify themselves with a unique hexadecimal code in the Master Boot Record at the beginning of each hard

disk drive. This unique hexadecimal code is commonly referred to as a 'Magic Marker' or 'Magic Byte' (Haider et al. 2012). However, forensic examination of Sector 0, the Master Boot Record, of the internal hard disk drive from the Ganz CCTV-DVR contained no Magic Marker and therefore did not contain one of the industry recognised file systems. In order to progress it would be a requirement to reverse engineer the unknown file system used on the internal hard drive. In order to do this you must first review how a file system stores data. In general before a file system can be created a partition is required to specify how much of the hard drive is to be used. These partitions equate out to a start point (cylinder) on the hard drive and an end point (cylinder). Once a partition is identified it is possible to establish the content of each area and how they relate to each over, building up a picture of communication as they progress.

Despite Master Boot Record area of the Ganz internal hard disk drive contains no magic marker to assist in the identification of the file system, there is other data which gave a starting point for research. The Master Boot Record of the Ganz internal hard disk drive was examined in X-Ways Forensic tool, it displayed the following (Figure 1):



Offset	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
0000000000	00	00	44	A9	4E	39	00	00	FB	FF	FF	39	00	00	00	FF	D©N9 ûÿÿ9 ÿ
0000000010	00	B0	4B	BA	01	00	00	00	00	00	58	6D	01	00	00	00	°K º Xm
0000000020	30	7E	F7	00	00	00	00	00	60	6F	F7	00	00	00	00	00	0~÷ `o÷
0000000030	55	41	56	54	45	43	48	AA	46	53	53	31	36	41	00	55	UAVTECH*FSS16A U
0000000040	00	3C	AD	BA	02	00	00	00	00	30	5E	38	3A	00	00	00	<-º 0^8:
0000000050	00	00	72	74	00	00	00	00	00	FF	AC	BA	02	00	00	00	rt ÿ¬º
0000000060	00	02	00	00	00	00	00	00	00	FF	71	74	00	00	00	00	ÿqt
0000000070	00	00	AD	BA	02	00	00	00	00	FF	AC	BA	02	00	00	00	-º ÿ¬º
0000000080	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	

Figure 1: Ganz internal hard disk drive – MBR (Sector 0)

The ASCII text AVTECH and FSS16A became relevant when it was established that the internal components of the Ganz DVR unit were manufactured with the brand AvTech. We conduct a search on both the branded company, Ganz and also the maker of the internal components AvTech. We realise that Ganz is a product brand of CBC America Corp. It would appear that the Ganz DVR unit model C-MPDVR-16 is an old product and as such there was very little information available on the unit. The only notable information is that the unit records with a MPEG format (http://en.cbc-cctv.com/uploads/tx_n21products/05_tab_088_01.gif).

Of particular importance to this paper is a thread on the CCTVforums.com website (CCTVforums.com), which discusses retrieving data from AvTech based units. This thread identified that AvTech devices use a proprietary file system which has not been decoded, however AvTech had released a program to assist in the interrogation of such DVR hard drives. The program is called Disk Tools.exe and a download link was provided. We also use this tool to compare forensic results with our approach.

Significantly the AvTech website also provides a warning that data may be destroyed if the DVR internal hard disk drive was connected to a PC and accessed. This would likely be due to Microsoft Windows trying to write a new recognised Master Boot Record to the hard disk drive on its initialization. In fact, our paper relates to a forensic examination with the use of a disk image created through a read only device. The actual Ganz DVR internal hard disk drive was never connected directly to any computer.

4. Forensic acquisition and analysis of Ganz CCTV: A Case study

4.1 Acquisition

The forensic process we used in this section is adopted from the Ariffin's model (Ariffin et al 2013). In our experiment, we perform forensic acquisition and analysis of a Ganz DVR Unit. The first step is identification where a Ganz C-MPDVR-16 DVR was photographed. We also take the photos at various stages of the examination. Figure 2 shows the internal view of this Ganz DVR Unit.

The next step is preservation. Within the law enforcement environment it is imperative that any data is recovered in a forensically-sound method. The general accepted criterion is that no changes are made to the original data source, and that any copies made are identical to the original data source. In our experiment, the Ganz DVR unit was seized from the working environment by law enforcement agents on the 4th of April 2013 shortly after 08:00hrs. A strict chain of custody from seizure to examination was in place and no access to the Ganz DVR Unit was permitted. For this examination the DVRs internal hard disk drive was removed and connected to a forensic device known as a write blocker. Next a bit copy was made of the hard disk drive. This

is a method to ensure that an exact copy of the entire hard disk drive is produced; including all used and used space.

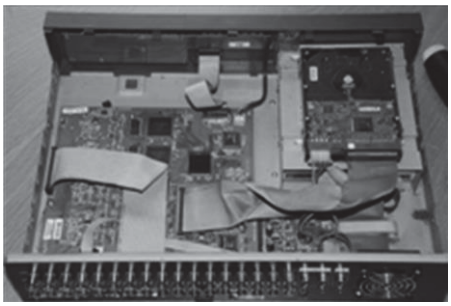


Figure 2: Internal view of the Ganz C-MPDVR-16 DVR unit

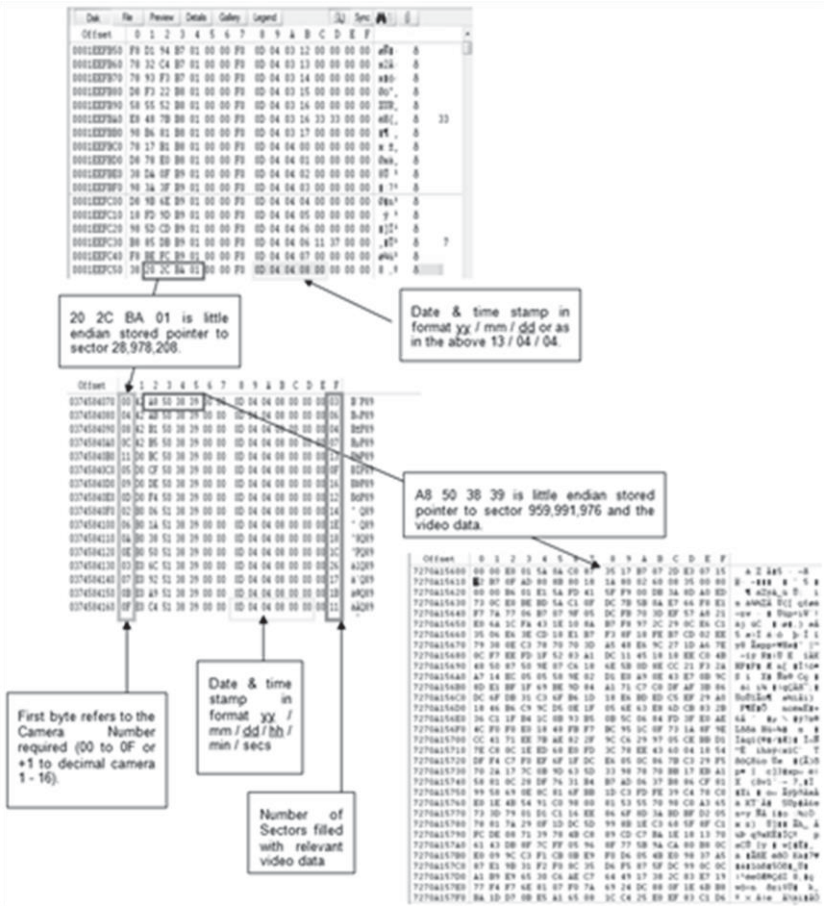


Figure 3: Overview of the AvTech proprietary file system

4.2 Analysis

The analysis took place on the clone hard disk drive, examining the information at the storage level to identify any recognised file signature. The key challenge with a proprietary file format is to locate the video streams without the assistance of predefined storage methods (FAT, NTFS etc.)

4.2.1 Initial actions

The cloned hard disk drive was opened within X-Ways Forensics, EnCase and WinHex none of which were able to recognise the file system or any video file. So, we examine the first sector of a hard disk drive (Master Boot Record). The only identifiable data in the sector 0 was the label UAvTech and FSS16A. No technical information was obtainable, other than FSS16A which is a proprietary file system of AvTech.

4.2.2 Identifying the data

Further examination was conducted to review the type of data visible. We found there are totally three distinct areas on the hard drive; the first contains entries which provide a general date and time for recorded footage and a pointer to the second area. The second area provides separate entries containing refined date and time stamps for each of the individual 16 cameras video footage recorded, and a pointer to the video footage. The third area contains the actual video footage (Figure 3). Due to the limitation of the paper size, we do not present in details the forensic process we used to locate these three areas. Besides, the main objective of our paper is to show how to recover and playback the recovered video data.

4.2.3 Retrieving data

In our experiment, one second period timeframe retrieval was attempted, 04/04/2013 08:00:00 hrs to 08:00:01 hrs. The data of this period is located in sector 28,978,208 / offset 0374584070 thru to sector 28,978,211 / offset 0374584720. By analysis of the headers in those sectors we notice that the video footage was captured at 6 frames per second. This carved data also revealed duplicate entries for cameras 0C, 0D, 0E, 0F on the 2nd and 5th frames, represented by bytes 1 of each entry showing FF. This pattern continued through all data and is possibly recording an error. In order to recover one second of video footage for camera 00 (6 frames per second) data required carving from the following sectors:

```
00 42 A8 50 38 39 00 00  0D 04 04 08 00 00 00 03
00 50 D5 51 38 39 00 00  0D 04 04 08 00 00 00 1A
00 51 8D 52 38 39 00 00  0D 04 04 08 00 00 00 0B
00 52 E0 52 38 39 00 00  0D 04 04 08 00 00 00 03
00 53 28 53 38 39 00 00  0D 04 04 08 00 00 00 03
00 54 80 53 38 39 00 00  0D 04 04 08 00 00 00 03
```

Sectors 959991976 +3
Sectors 959992277 +26
Sectors 959992461 +5
Sectors 959992544 +3
Sectors 959992616 +3
Sectors 959992704 +3

All data was carved and compiled into one file; it totalled 21.5k in size.

4.2.4 Carved data identification

In order to identify the data within the carved file, we tried a number of industry standard applications including MediaInfo 0.7.69, GSpot v2.70a, VideoInspector 2.6.0.129. However, none of the above software was able to detect the video format contained in the carved file, indicating that it was likely a proprietary encoding.

The Ganz DVR system was supplied with proprietary viewing software called Video Player MFC, version 1.1.6.1.

The application stated it played the following proprietary file formats: .VS4, .VSE, .DVR, .AVC, .DV4. So, we made five copies of the carved data file and each provided with one of the five file extensions above. These were then loaded into Video Player MFC. The only successful access within the video player was the copy of the carved data given the .DVR extension. Thus the video player identified the clip as being from the 04th April 2013 at 0800hrs as expected, it also displayed that the footage was recorded in the 720 x 576 pixel format (Figure 4).

Based on our investigation, the Ganz DVR is listed as recording in MPEG4 format. A review of the carved file used previously showed that each segment of video data started with the hex: 00 00 E0 01. According to the above MPEG file header format (http://mpeg.chiariglione.org/) a standard MPEG file would start with: 00 00 01 ??. With the ?? being replaced by the relevant bit above, which for a video stream would be E0 to EF depending on the stream. So, if we place the two sections of hex codes together there is a clear similarity:

Carved data file - 00 00 00 E0 01
MPEG Format - 00 00 00 01 E0 (first video stream)

It was possible that the Ganz DVR simply transposed bytes 3 and 4. In order to test this theory, we used a hex editor to modify a copy of the carved data file swapping bytes 3 and 4 in each of the stream headers. This modified data was saved as carved.mpg file and attempted to be opened in a number of media players. However the carved.mpg file would not play. The carved.mpg was further analysed with the previously referenced application MediaInfo. On this occasion MediaInfo recognised the carved.mpg file as a MPEG-PS stream, but it refused to provide any details of codec used for encoding. Indeed, it appeared that the Ganz DVR uses a proprietary codec for encoding and storing the video streams in the DVR file type. Further research was required

into the actual data streams, it was elected to analyse the first 32 bytes of the start of each video stream for two specific channels (channel 1 & 2) at a specific time (08:00hrs) over a 4 day period spanning a month change to provide a reference guide to decoding the file header.

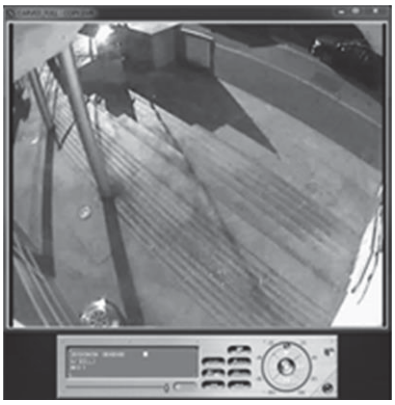


Figure 4: Video player MFC showing carved data file

4.2.5 Video file header investigation

In this experiment, we extract the first 32 bytes of the video stream for channel 1 & 2 at 0800hrs on the 30/03/2013, 31/03/2013, 01/04/2013 and 02/04/2013 (Figure 5). In order to establish whether offsets 30 and 31 (the last two bytes: 00 80) provide a time stamp, we retrieve data from the camera 1 on the 30th of March 2013 for each hour recorded between 0800hrs on the 30th until 0000hrs on the 31st. We then find that Offset 30 and 31 was linked to the time, it was clear that the Ganz DVR recorded the hour directly until it reached 1600hrs. At that point it resets offset 31 to 00 but increased offset 29 by 1. In this first test, we only focus on hours: 08:00, 09:00, 10:00hrs etc. There was no provision for minutes or seconds to be accounted. In order to locate the minute change additional data was retrieved splitting out the 08:00hrs timeframe for the 30th of March 2013. In reviewing the results we notices that offset 31 did not remain at 80 as expected if it was to indicate 08:00hrs. Neither was offset 30 remaining constant with the time stamp for minutes. Rather offset 30 appeared to initially be recording seconds within the first minute 08:00 – 08:01hrs. This appeared to be in direct hex notation, i.e. 00 – 3B (Dec 00 to 59). However at the beginning of 08:01hrs, offset 30 became hex 40. Then at the beginning of 08:02hrs, offset 30 became hex 80.Then at the beginning of 08:03hrs, offset 30 returned to hex C0. Finally at 08:04hrs offset 30 reset to 00. It was noted that offset 31 increased by 1, from 80 to 81. In visual form offset 30 represented:

Hex 00 – 3B = 0 to 59 Sec, Minute 0
Hex 40 – 7B = 0 to 59 Sec, Minute 1
Hex 80 – BB = 0 to 59 Sec, Minute 2
Hex C0 – FB = 0 to 59 Sec, Minute 3

The analysing of headers is shown in Figure 6. The difference between the starts of each cycle was noted as decimal 64. (64, 128, 192). The value in seconds could be established from the hex code in the following manner:

Example: offset 30 = hex FB = Dec 251
Dec 251 / 64 (cycle) = 192 r 59
Dec 192 / 64 = 3
Therefore Hex FB = 3rd Minute cycle, 59 seconds

Offset 31 was clearly recording two separate portions of data. Bit 1 was the direct hex value of the hour. Bit 2 was a record keeper of how many cycles offset 30 had completed within that hour.

Example: offset 31 = hex 8E
Bit 1 reads hex 8 = dec 8, therefore 0800hrs. Bit 2 reads hex E = dec 14, there 14 x 4 minute cycles = 56
Time stamp would read: 08:56hrs + value from offset 31
30th March 2013 – 0800hrs

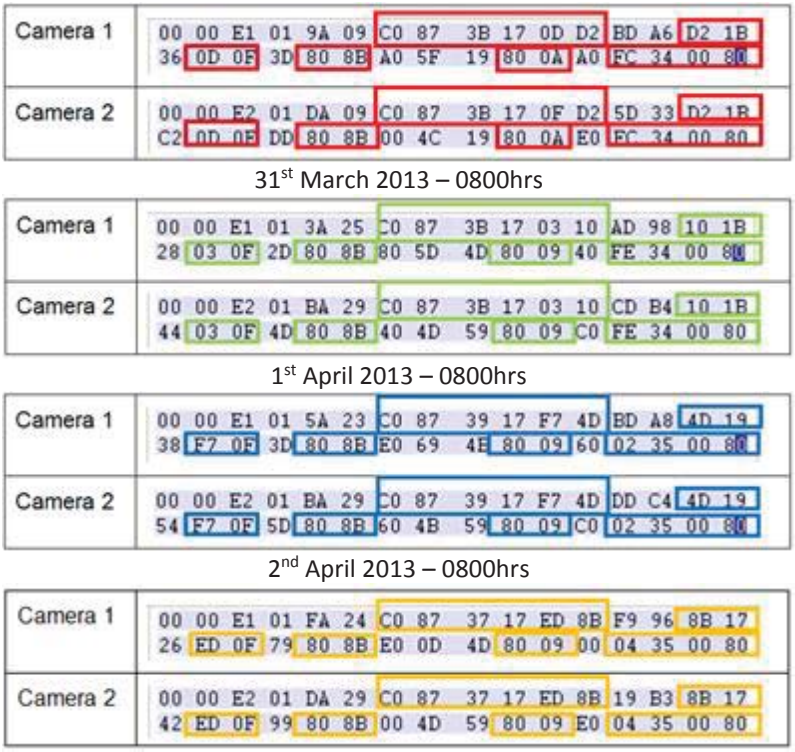


Figure 5: The first 32 bytes of the video stream for channel 1 & 2

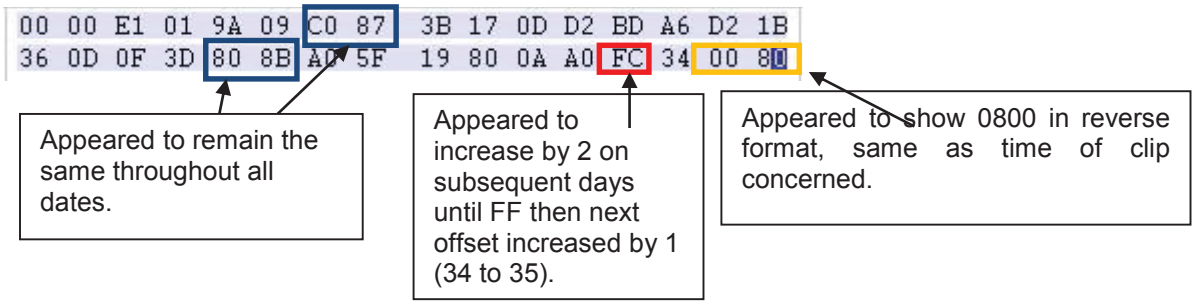


Figure 6: Video stream – header analysis

Next, looking at an overall example where we extract a video header from Sector 696,662,102 (Figure 7). If the above is to be proven then the video footage in Sector 696,662,102 should relate to footage taken at 08:32:42hrs. So we reverse the lookup method for sectors from date. We have sector 696,662,102 = Hex 29 86 38 56, transposed into little endian for search = 56 38 86 29 produced (Figure 8).

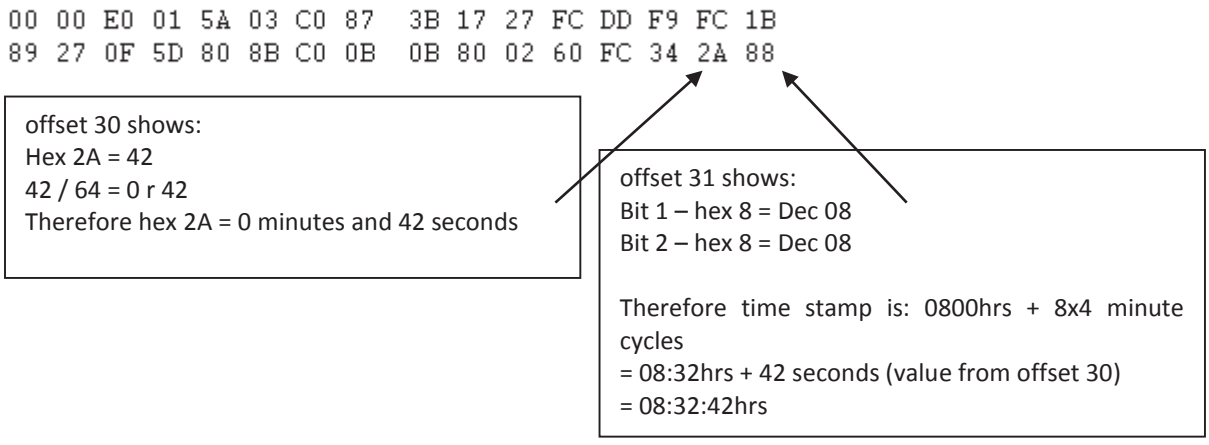


Figure 7: Video header from sector 696,662,102

Figure 8: Footage analysis

This result established that the footage in sector 696,662,102 was for the 30/03/2013 at 08:32:42hrs as per the example workings above and demonstrates the time function has been correctly deciphered. Next, we are looking at the date stamp artefacts. Based on the time stamp analysis it was evident that the date stamp was somehow linked with offsets 28 and 29, and that 16:00hrs played a crucial role in increasing the offset 28 counter. In order to establish the system used for the date stamp, we take the video footage from Camera 1 at 0000hrs and 1600hrs on the 30th March, 31st March, 1st of April and 2nd of April. This selection would provide four day changes and one month change. (Figure 9)

30th March 2013 - 0000hrs	00	00	E0	01	3A	1D	C0	87	37	17	13	68	B9	B0	68	17
	40	13	0F	3D	80	8B	40	4A	3D	80	05	40	FC	34	00	00
30th March 2013 - 1600hrs	00	00	E0	01	3A	36	C0	87	31	17	0B	3C	09	97	3C	11
	26	0B	0F	89	80	8B	20	6D	6E	80	01	40	FD	34	00	00
31st March 2013 - 0000hrs	00	00	E0	01	7A	0A	C0	87	35	17	07	A6	35	C2	A6	15
	51	07	0F	B5	80	8B	00	4A	1A	80	02	80	FE	34	00	00
31st March 2013 – 1600hrs	00	00	E0	01	DA	35	C0	87	3F	17	FD	79	9D	C6	79	1F
	56	FD	0F	1D	80	8B	20	8B	6D	80	01	E0	FF	34	00	00
1st April 2013 – 0000hrs	00	00	E0	01	1A	09	C0	87	33	17	F9	E3	91	F1	E3	13
	81	F9	0F	11	80	8B	C0	0F	19	80	0A	20	02	35	00	00
1st April 2013 – 1600hrs	00	00	E0	01	3A	0A	C0	87	3D	17	F7	B7	D1	05	B7	1D
	95	F5	0F	51	80	8B	20	1B	1A	80	02	40	03	35	00	00
2nd April 2013 – 0000hrs	00	00	E0	01	9A	1D	C0	87	33	17	F1	21	91	C0	21	13
	50	F1	0F	11	80	8B	60	4A	3D	80	05	A0	04	35	00	00
2nd April 2013 – 1600hrs	00	00	E0	01	1A	06	C0	87	3B	17	EB	F5	8D	84	F5	1B
	14	EB	0F	0D	80	8B	60	0E	0E	80	02	20	05	35	00	00

Figure 9: Hex dump of **v**ideo footage of four days

We notice that the hexadecimal value present in offset 28 increases at 0000hrs and 1600hrs of each day. However the data of offsets 28, 29 and 30 do not appear to match any of the industry date stamp formats (UNIX, MS-DOS etc.) and is likely a time counter from some point set by the manufacturer. Further examination would be required to establish the date stamp encoding method.

4.3 Proprietary application vs forensic examination

In this section, we use DiskTools.exe to examine the Ganz DVR unit internal hard disk drive to compare with our forensic results. The DiskTools.exe application stated that the earliest available video footage recoverable was from the 18th of March 2013 at 00:48:09 hrs. (Figure 10)

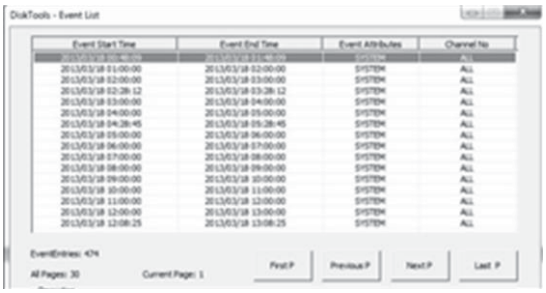


Figure 10: DiskTools.exe

So our forensic analysis was conducted for the date / time stamp 0D 03 12 00 30 09. The first reference appeared at Sector 63,343 which pointed to Sector 23,943,068 and finally to Sector 973,078,543 where the video data was located. The hex code for the time stamp was equated as:

0000hrs + (12x4)mins + 9secs = 00:48:09 hrs, which was identical to the timestamp expected. The footage suggested by DiskTools.exe was verified. In order to test the DiskTools.exe accuracy the previous timestamp from Sector 63,343 was taken and the pointers followed:

```
F8 95 31 6D 01 00 00 F0  0D 03 12 00 00 00 00 00
```

This pointed to Sector 23,933,333:

```
0C FF 8F EB E4 39 00 00  0D 03 12 00 00 00 00 05
```

This pointed to Sector 971,303,982

```
00 00 E0 01 1A 1F C0 87  39 17 BB 85 49 F3 85 19
82 BB 0F C9 80 8B 00 49  3F 80 01 20 E4 34 00 00
```

This had video footage present, with a timestamp reflecting 00:00hrs. In order to view whether the footage was the correct footage the data was extracted and converted using the methods within this paper. The sectors concerned were: Sector 971303982 - (10 sectors), Sector 971304138 - (6 sectors), Sector 971304220 - (6 sectors), Sector 971304320 - (5 sectors), Sector 971304398 - (5 sectors) and Sector 971304475 - (5 sectors).

The carved data was saved to the file 0000hrs.DVR and opened in the Video Player.

Figure 11: Carved data playback

The Video Player displayed footage from the 18th March 2013 at 00:00hrs. This footage was not retrievable through the DiskTools.exe application. In order to establish the oldest video footage available a trial and error system was deployed, starting with the attempt to retrieve footage from 2300hrs on the 17th March 2013.

Sector 23921205 contained the pointer for 0D 03 11 17 00 00 or in normal terms 17th March 2013 at 2300hrs. Sector 23921205 pointed to Sector 969083922. Sector 969083922 contained video data with the header of:

```
00 00 E0 01 FA 0A C0 87  39 17 7D 38 0D 4A 38 19
D9 7B 0F 8D 80 8B 40 4A  1B 80 0A 00 E3 34 00 70
```

This header indicates that the footage is for 23:00hrs (00 70 with the +16hrs from E3). The next trial and error search was conducted for the 16th of March 2013 at 2300hrs. Sector 23630109 contained the pointer for 0D 03 10 17 00 00 and pointed towards Section 916459161. However Sector 916459161 contained video data without a header:

```
E6 05 F0 00 3B 2F 35 DE  82 81 8C 1E 4D 79 DC BB
07 1E 35 5D B7 DB EB 22  06 1C 38 24 CC 0C 3E 74
```

This data is clearly not a start of a video stream as the 00 00 E0 01 starting bytes are missing. Further reading showed that Sector 916459161 contained video data from a stream that started in Sector 916459157, which contained a timestamp for 12:53:51 hrs. Continuing the trial and error method established that the oldest footage available was from 20:00hrs on the 17th of March 2013. Therefore an additional 4hrs and 48 minutes of footage was available through manual examination of the Ganz DVR hard drive, compared to the official DiskTools.exe application. In criminal investigations any additional time recoverable may result in crucial evidence, which if reliant on the official application would not have been recovered. Why the official tool missed these 4hrs and 48 minutes was concerning so further investigation was conducted.

5. Conclusion and future work

In this paper we proposed a new method of reverse engineering the proprietary file system of a CCTV Digital Video Recorder. By conducting the examination as shown in this paper the file system has been reversed engineered with no initial knowledge, applications or directions available. Further it has been reversed engineered to a sufficient degree to allow for the identification and retrieval of video footage from any specified camera for any specified date recorded without the use of any proprietary applications. Whilst this paper was unable to evidence how to decode the date stamp from within the actual video footage data, this is a redundant step as the date and time stamps are available from the first two locators as referenced in Section 4. Further research into the actual video data stream may reveal how the date stamp is recorded, this would assist when presented with a raw partial data stream was provided for examination i.e. no locator data is available due to damaged hard drive etc. Overall the results are directly transferable to any CCTV Digital Video Recorder system that uses the AvTech file system. With minor alterations the process contained within this paper can be utilised with any proprietary file system. We are also looking at combining similar approaches in Vehicle Forensics (Jacobs 2016) and Mobile Device Forensics (Faheem 2015, Sgaras 2015) to improve our method.

References

Ariffin, A., Slay, J., Choo, K-K. (2013), Data Recovery from Proprietary Formatted CCTV Hard Disks Digital Forensics, Chapter in Advances in Digital Forensics IX, Volume 410 of the series IFIP Advances in Information and Communication Technology pp. 213-223

CCTVforums.com http://www.cctvforum.com/viewtopic.php?f=56&t=24717

Dongen, W. S. V (2008) Case Study: Forensic Analysis of a Samsung digital video recorder, Journal of Digital Investigation, vol. 5, pp. 19-28, 2008.

Faheem (2015) Faheem, M., Kechadi M., Le-Khac, N-A., The State of the Art Forensic Techniques in Mobile Cloud Environment: A Survey, Challenges and Current Trends, International Journal of Digital Crime and Forensics (IJDCF), Vol 7(2) p.1-19

Haider, Dr. , Al-Khateeb M., (2012) Analyzing the Master Boot Record, Webspace, 2012

Han, J., Jeong, D. and Lee, S. (2015) Analysis of the HIKVISION DVR File System, Digital Forensics and Cyber Crime: 7th International Conference, ICDF2C 2015

IMS Research (2012), Trends for 2012

Jacobs (2016) Jacobs, D., Le-Khac. N-A., Vehicle Entertainment System Forensics: A Case Study of Volkswagen Automobile, Twelfth Annual IFIP WG 11.9 International Conference on Digital Forensics, New Delhi, India, January 2016

McKemmish R. (1999) What is forensic computing?Trends & Issues in Crime and Criminal Justice 1999;118:1-6.

Poole, N.R., Zhou, Q. and Abatis, P. (2008), Analysis of CCTV digital video recorder hard disk storage system, Digital Investigation, vol.5, no.1, pp. 85-92, May 2008.

Sgaras (2015), Sgaras C., Kechadi, M-T., Le-Khac, N-A. Forensics Acquisition and Analysis of Instant Messaging and VoIP Applications, Computational Forensics, Springer International Publishing, 2015 p.188-199

Tobin, L., Shosha, A., Gladyshev, P., (2014) Reverse engineering a CCTV system, a case study, Digital Investigation vol.11(3) pp. 179-186

Wang, W. (2009), Digital Video Forensics, PhD. Thesis, Darthmouth College, New Hampshire, USA, June 2009

Zeltser, L. (2001) “Reverse Engineering Malware” https://zeltser.com/reverse-engineering-malware-methodology/

Intrusion Detection in Cyber Physical Systems Based on Process Modelling

Tamás Holczer, András Gazdag and György Miru
Laboratory of Cryptography and System Security, Budapest University of Technology and Economics, Hungary
holczer@crysys.hu
agazdag@crysys.hu
miru@crysys.hu

Abstract: Cyber physical systems (CPS) are used to control chemical processes, and can be found in manufacturing, civil infrastructure, energy industry, transportation and in many more places. There is one common characteristic in these areas, their operation is critical as a malfunction can potential be life-threatening. In the past, an attack against the cyber part of the systems can lead to physical consequences. The first well known attack against a CPS was Stuxnet in 2010. It is challenging to develop countermeasures in this field without endangering the normal operation of the underlying system. In our research, our goal was to detect attacks without interfering with the cyber physical systems in any way. This can be realized by an anomaly detection system using passive network monitoring. Our approach is based on analysing the state of the physical process by interpreting the communication between the control system and the supervisory system. This state can be compared to a model based prediction of the system, which can serve as a solid base for intrusion detection. In order to realize our intrusion detection system, a testbed was built based on widely used Siemens PLCs. Our implementation consists of three main parts. The first task is to understand the network communication in order to gain information about the controlled process. This was realized by analysing and deeply understanding the publicly undocumented Siemens management protocol. The resulting protocol parser was integrated into the widely-used Bro network security monitoring framework. Gathering information about the process state for a prolonged time creates time series. With these time series, as the second step, statistical models of the physical process can be built to predict future states. As the final step, the new states of the physical process can be compared with the predicted states. Significant differences can be considered as an indicator of compromise.

Keywords: intrusion detection, process modelling, cyber physical system, anomaly detection

1. Introduction and background

Industrial control systems (ICS) are special cases of cyber physical systems (CPS). The security of CPSs is an important issue especially if it is an ICS. ICSs can control factories, electric grids or many other important systems. It is common in these systems, that an attack can have tremendous effects.

It is hard to develop defence mechanisms in this field without endangering the normal operation of the underlying system. For this reason, it was decided to implement a passive detection system, which cannot interfere with the normal operation of the controlled process, but can trigger an alarm, in case of suspicious events.

In order to realize our intrusion detection system, a testbed was built based on widely used Siemens PLCs. Our implementation consists of three main parts. The first task is to understand the network communication in order to gain information about the controlled process. This was realized by analysing and deeply understanding the publicly undocumented Siemens management protocol. The resulting protocol parser was integrated into the widely-used Bro network security monitoring framework. Gathering information about the process state for a prolonged time creates time series. With these time series, as the second step, we can build statistical models of the physical process to predict future states. As the final step, the new states of the physical process can be compared with the predicted states. Significant differences can be considered as an indicator of compromise.

The following attacker model was used throughout our work. The attacker has direct access to the physical process. It can be a physical or logical access as well, but it modifies the measured variables of the process. In most cases, it is implemented by modifying the process and masquerading the process output from the operator by modifying the output of the human machine interface (HMI). It was assumed that the real sensor readings could be grabbed from the network communication. This kind of attack can be detected by this approach. In this attacker model, it is assumed that there is a clear learning phase before any attack is done.

In the next sections it is shown how the system used differs from the state of the art solutions, how the system was built, and finally how accurately it could detect attacks.

2. State of the art

In conventional IT intrusion detection has a rich tradition. Axelsson *et al*. (2000) reviews and categorizes the existing IDS solutions based on the used detection principle and certain operational aspects of the system. Debar *et al.* (1999) further refines the taxonomy of intrusion detection systems and defines families according to their properties.

Intrusion detection in industrial control systems is a relatively new field of study, however there are many notable articles on the subject. Zhu *et al.* (2010) investigates the SCADA specific intrusion detection techniques and systems, provides a definition, taxonomy and a set of metrics to compare existing solutions. The paper also tries to ease interoperability between traditional IT security and ICS research by providing cross discipline insight to the detection problems of both fields.

Valli (2009) outlines an approach of using traditional IT applications (Snort, nepenthes, honeyd) to create a resilient layered defence application for control system networks. Verba *et al.* (2008) proposes SCADA specific anomaly and signature based packet inspection mechanism and traffic flow analysis with fine protocol granularity. Garitano et al. (2011) reviews research done in the field of anomaly based intrusion detection in ICS applications and evaluates the viability of such methods. Yang *et al.* (2006) describes an anomaly based IDS for control systems, the implemented method uses auto associative kernel regression model with statistical probability ratio test to detect anomalies in the packet flow. The proposed system is applied on a simulated real time SCADA system and the test results are evaluated.

Kiss *et al.* (2014) uses data clustering and Big Data technologies for real-time analysis of control network and sensor data in order to identify physical threats. Hadžiosmanović *et al.* (2014) introduces a system to detect anomalies in the monitored process variables. The proposed method extracts the data from the network and classifies the variables into three different groups. He identifies constant, discrete in a domain and dynamic variables and uses auto regression to detect anomalies.

3. System design

Our system consists of three main parts. These main parts are shown on the following figure.

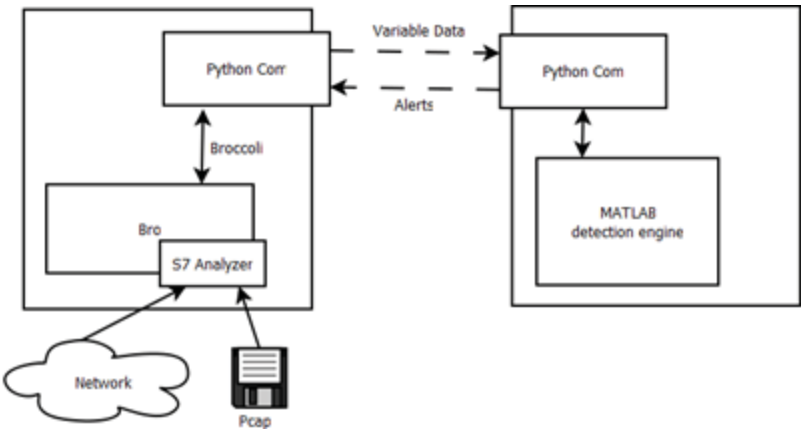


Figure 1: System overview

The first part is responsible for intercepting the network traffic and analysing it. This part is realized by a newly developed Bro analyser. The second part is responsible for the communication between the Bro analyser and the detection engine. This part is implemented by some Python scripts. The third part is responsible for the anomaly detection. This part is implemented in Matlab. In the following sections, these parts are introduced, before evaluating the whole system.

3.1 Message interception

As discussed in the previous sections, the general design of the proposed IDS requires the process data to be extracted from the network communication. As a result, the proposed system needs to be able to observe and understand the communication between the field devices and the entities managing them. Capturing the traffic of large, distributed IP networks is a nontrivial task. However there are multiple existing solutions and papers covering the subject. In the following discussion we assume that the entire network traffic is available in a capture file or monitored to a single interface of a dedicated PC.

The IDS aims to protect ICS systems built with specific Siemens equipment that uses the proprietary S7 protocol. This is a closed, undocumented protocol. The reverse engineering of the protocol was required to an extent to be able to extract valuable process related information from the communication. The main technical challenges we faced were to get an understanding of the S7 protocol and create a software piece that can parse it. Also, the software needed to read variable data that can be passed to the detection engine. The rest of this segment shortly describes the reverse engineering process and details the development of the protocol parser.

The Siemens S7 protocol is widely used in the industry and it has been the interest of many to acquire a deep knowledge of the protocol. Although, at the time of this writing no comprehensive documentation exists there are notable projects that need to be mentioned. Davide Nardella has created an open source communication library the Snap7 (Nardella, 2015), which implements basic communication scenarios. The library comes with the extensive documentation of the basic structure of the S7 protocol. Another project is the S7 Wireshark dissector by Thomas W. which covers most of the protocol and its source code and contains a lengthy list of protocol constants. None of these projects are complete, yet they are invaluable when dealing with S7 communication.

The Siemens TIA portal and WinCC Advanced were used as communication masters and an S-300 series PLC as the slave device, and observed the communication between them in Wireshark network analyzer. The main challenge we encountered was that the S7 protocol turned out to be bloated and redundant in functionality. The protocol contains multiple addressing modes and multiple ways to execute the same actions. We had to uncover each of these in order to not miss any valuable data exchange between the entities.

The software analyzing the protocol is required to process live network streams or capture files, extract the needed variable data and send it the Matlab detection engine. Also, there are secondary requirements such as logging raw data and detection events and presenting alerts in case an intrusion is suspected. To implement these functionalities multiple approaches have been considered, these are the following:

- Create a custom program from scratch
- Use an existing scriptable network analyzer
- Use an existing and extensible IDS implementation

All these methods have their pros and cons, a custom program provides the most flexibility, however it might require considerably more effort to implement, integrate and test. Building the protocol analyzer on top of an existing IDS provides the benefit of having a tested and established code base fulfilling all the secondary requirements, thus greatly reducing the amount of the implementation work. It was decided that the additional benefit of an existing IDS outshines the flexibility provided by the custom implementation.

Multiple major IDS programs were considered when deciding on the base software. The Bro IDS for the S7 protocol parsing was chosen because it is a well-established, versatile, scriptable detection framework. In addition, it provides means through communication APIs to pass the extracted data to further processing which was a crucial requirement for the detector. It also appeared to be more flexible and customizable than the other candidates.

The Bro inner structure follows the conventional actor model, where each actor can publish or subscribe to certain events. Bro provides a domain specific scripting language that is used to define event handlers or publish new events, create data structures, write arbitrary program code and use the services of existing frameworks. These frameworks provide a variety of functionality such as logging, creating alarms and notices, geolocating IP addresses and offer signature based detection.

The performance critical parts of the Bro IDS are written in C++, and there is an interoperability layer between the native and the scripted stack. The Bro framework is extensible with plugins which can contain native code and bro scripts as well, they can implement arbitrary functionality. One use of such plugins is to write protocol analyzers for communication protocols not yet implemented in Bro. Inside the program, these analyzers are used to parse the different network protocols, extract the required information and raise events that are later handled in the scripted layer. Unfortunately, this part of the bro development is not yet documented; however there are plenty of different analyzer examples in the Bro sources and plugins can access the same APIs and classes as any source code.

We have decided to implement the protocol parser in such plugin. The analyzer receives the S7 packets and the connection state information, which allows it to store data between calls and map specific requests to their replies. This is essential when extracting the process variables, due to the way memory reading is implemented in the S7 protocol. During these memory reads only the request sent by the master contains the variable address and size information the slave replies with the raw data bytes. The variable data extracted by the analyzer is sent to the upper layers for further processing. The plugin also parses other, non-process-variable-related S7 communication events, which can be used for conventional rule based intrusion detection. Such events are: program block modification, diagnostic data read, PLC control commands, firmware update, read/write request to certain memory areas, password authentication.

The data and connection events are processed by upper layer bro scripts and logged by the logging framework. The Bro IDS is equipped with the Bro Communication Client Library (BroCCoLi), the library can be used to send to or receive events from other programs. It is written in C, however it has python and ruby bindings. The S7 data handler bro script sends the extracted data to the python script which calls the Matlab detection engine and sends back detection events if there is any. The data handler script can raise alerts and notices based on these detection events. The event flow inside and outside Bro is presented in the picture below.

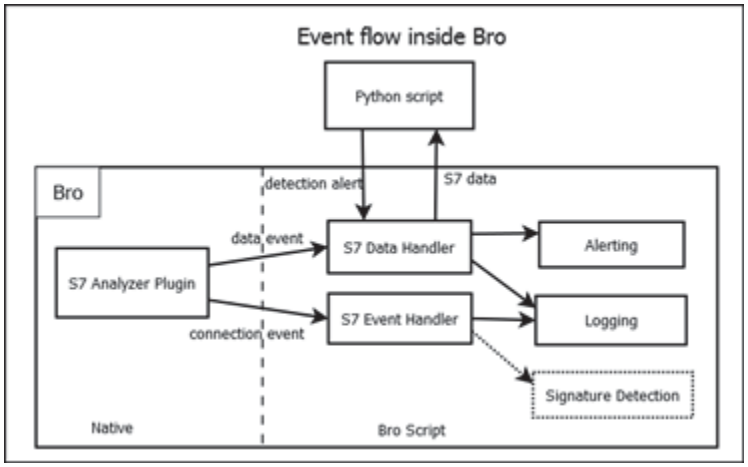


Figure 2: Event flow inside bro

3.2 Middleware

The task to solve for the middleware was to establish communication between the Bro part of the system and the Matlab engine. This problem had to be solved within the boundaries of various further prerequisite.

The Bro engine runs on Linux, whereas our Matlab version runs on Windows and has a number of interfaces for communication. We have chosen the python language to build the bridge between the systems.

On the Bro side an inter process communication had to be used to establish a connection. Fortunately, Bro comes with support for that called the BroCCoLi library. This library lets the developer call an external python function to export Bro data. It works in the other way as well, allowing initiating Bro events from the outside. The remaining task to be solved was to convert the internal data types and structures from Bro to standard python types. This was required to be able to send the data through the network connection.

On the Matlab part another inter process communication realization was required. Matlab has a plugin to work with python scripts. This wraps the Matlab engine into a python object allowing the user to execute Matlab functions from within the python context. A further task to solve was to interpret the data for the Matlab engine and then evaluate the result.

Based on the Matlab engine decision a returning network communication may be required to alert the Bro engine, and with that the network operator as well, about a potential attack.

3.3 Model based detection

In intrusion detection systems, many approaches exist to detect malicious activities. The two main approaches are the signature based detection and the anomaly based detection. Signature based detection uses previously known intrusion signatures, while anomaly detection detects the changed behaviour of the system. In this work we used anomaly detection based approach, because the attack signatures are not previously known in industrial control systems.

Many anomaly detection based approaches exist in the literature, but the model based approach suits best our purposes. Its main idea is to build a model of the investigated process, and detect if the model and the real measurements are diverging. If the difference between the model and the real measurements is over a threshold, an alarm can be raised. In our case the model is built using clean measurements, when no attack is assumed in the system. The main building blocks are shown on the following figure.

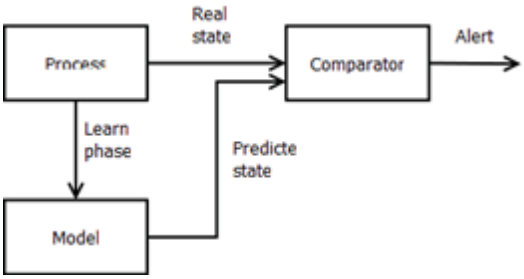


Figure 3: Model based detection

The advantages of model based detection are that it can detect unknown attacks, and the model adapts itself automatically to previously unknown processes. The main disadvantages of model based detection are false alarms and the need for clear traffic.

In the following section, we will show how the model based detection works in industrial control.

4. Evaluation

In this section we will show how our model based intrusion detection systems works in industrial control system environments. First two different approaches are compared. They are the autoregressive–moving-average (ARMA) models and neural networks. The two models are analysed, and the problem of parameter selection is covered as well. After finding the more promising model and parameters, the model is evaluated in a different environment.

4.1 Comparison of models

The first model we used is the autoregressive–moving-average (ARMA) model. It models a weakly stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average:

$$X_T = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1}$$

The other model candidate we used is nonlinear autoregressive neural networks. They can be trained to predict a time series from that series past values.

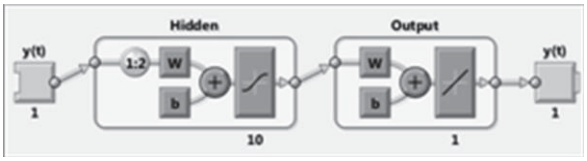


Figure 4: NARNET model

Both models were analysed by a simple process. The process contained a tank, where randomly changing amount of water is filled in. The control process kept the level of water between a minimum and a maximum value by opening and closing a valve. The schematic of the process is shown on the following figure.

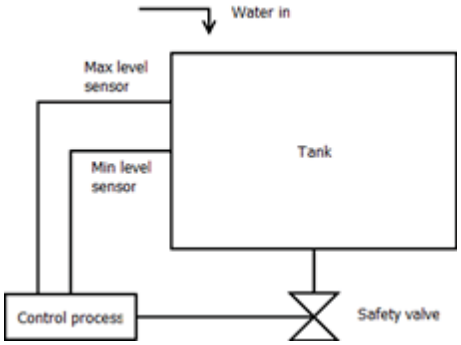


Figure 5: Test process

When deciding between the two models and the parameters, the average accuracy of the model, the standard deviation of the model and required time of running was analysed based on Matlab implementations of the models. The results are shown on the following figures. The ARMA model was analysed with parameters running from 1 to 40 (degree of the polynomials). The parameters of the NARNET model were the number of hidden layers and the delay of the feedback. Every experiment was run hundred times. The value of the fit is the mean value of the hundred runs.

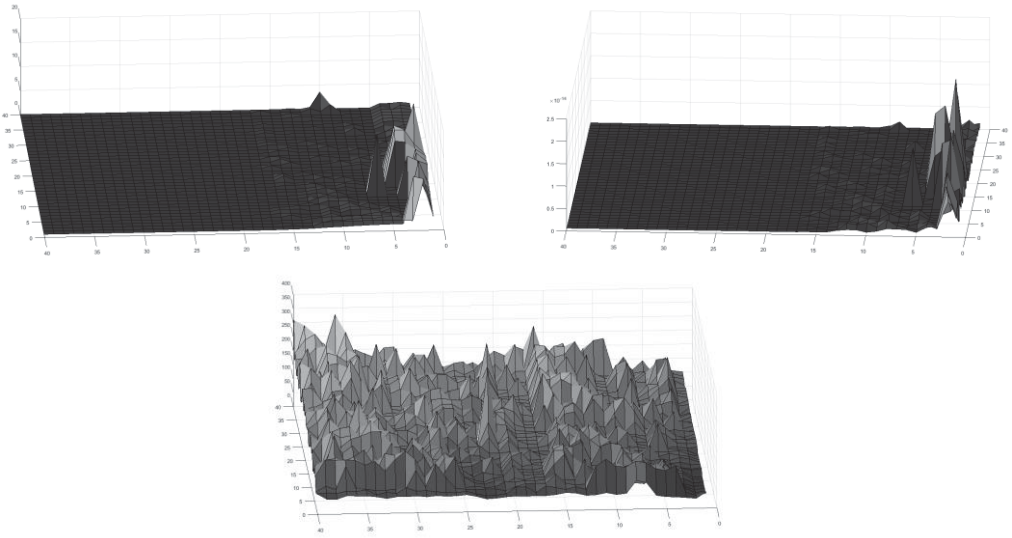


Figure 6: ARMA fit, standard deviation and run time

It can be seen that both models works well in terms of predicting the next value of the process, however the ARMA model is better. In terms of standard deviation, the ARMA model is highly superior to the NARNET model, which means that the results from the ARMA model are more reliable. In terms of run time, there is no significant difference between the two models. As a conclusion we decided to use the ARMA model, because it is more accurate and more reliable compared to the NARNET model. In terms of parameters, the 20 degree polynomials were selected, as they are accurate, reliable, and the run time is acceptable.

In the next section we analyse how the selected ARMA model behaves with different processes.

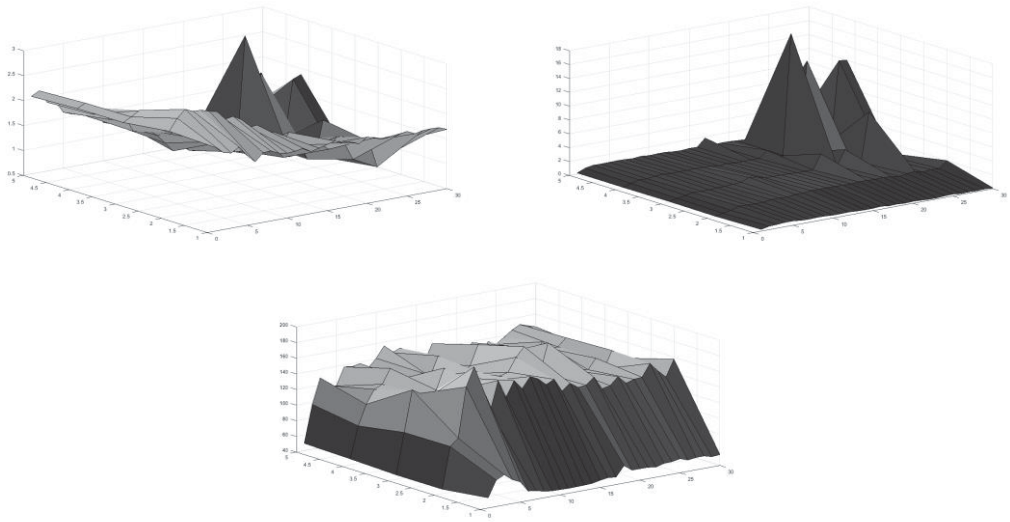


Figure 7: NARNET fit, standard deviation and run time

4.2 Analysis of the ARMA model

We evaluated our chosen Modell with 4 different tests. We tested in both normal operation and in an attack scenario the robustness of the model with different teaching and testing data length. For this test we used the Tennessee Eastman Process (Downs and Vogel, 1993).

The Tennessee Eastman Process model of Down and Vogel (Downs and Vogel, 1993) remains to be one of the most important tool for evaluating system theory concepts and validating algorithms. It is based on a real chemical process that causes the model to be a complex multicomponent system. Due to the wide usages of this process it has multiple implementations in a number of languages. For our test purposes we used a revised Matlab implementation from 2015 by Bathelt *et al* (2015).

The process consists of a reactor/separator/recycle arrangement containing two simultaneous gas-liquid exothermic reactions of the following form:

A(g) + C(g) + D(g) -> G(liq), Product 1,

A(g) + C(g) + E(g) -> H(liq), Product 2.

A sematic figure of the process looks like the following (Bathelt *et al*, 2015):

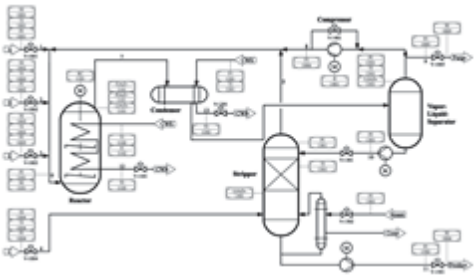


Figure 8: Tennessee Eastman process

This implementation has numerous output parameters that could have been used for model evaluation. We have chosen to use the reactor pressure parameter because of its intuitive interpretation. During multiple runs of the process pressure values were collected and ordered them into an array for the test.

The four conducted test were the following. The model was tested with different teaching intervals to measure its learning quality. After that the model was tested with different evaluation periods to test its predication quality as well.

The other two tests were to evaluate the attack detection mechanism. This time we performed the same two tests again but with known attack vectors. The results of our test are the following:

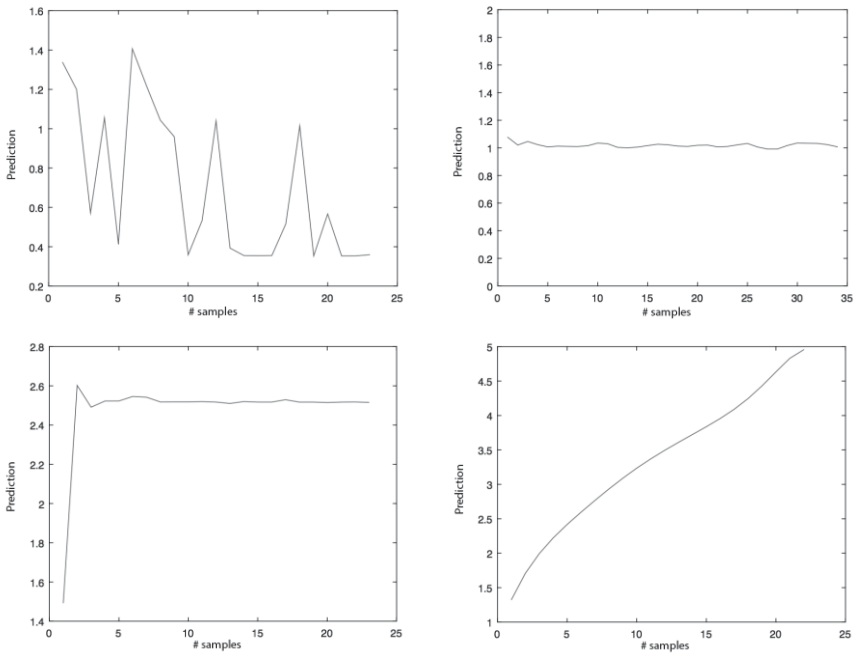


Figure 9: ARMA model analysis results

The first two figures show the results for our teaching tests. In case of a clean system, with the increase of the teaching set, the prediction gets better. It is clear that the more the model knows about the system, the better it can predict the future output. If we have at least 20 clear samples, the error of the prediction gets reasonably low.

Once we have a proper model, the increase of the prediction length makes no significant difference for the evaluation (as shown in the second figure). The ARMA model assigns the same value (indicating no discrepancy) continuously to the system.

The second set of images show the system results in case of an attack. With an increase of the teaching set the prediction results gets better. A high prediction error means that the predicted and measured values differ in a significant way. The test shows that if the clean set is long enough (at least 3-4 samples), we can detect the attack successfully.

On the last figure, our final test shows that an increase of the prediction length increases confidence of an attack prediction. This behaviour is in correspondence with our exceptions.

5. Summary

In this paper, it is shown how an intrusion detection system, which can detect attacks against industrial control systems using programmable logic controllers, was built. This solution is superior to previous approaches in terms of accuracy and usability.

According to our results, our system can detect different intrusions by identifying anomalies in the state transitions of the physical process. This approach can be used in any critical CPS, but it can be especially valuable in legacy systems, where additional security can be achieved without any modification of the system.

References

Axelsson, Stefan. Intrusion detection systems: A survey and taxonomy. Vol. 99. Chalmers University of Technology, Goteborg, Sweden: Technical report, 2000.

Bathelt, Andreas, N. Lawrence Ricker, and Mohieddine Jelali. "Revision of the Tennessee Eastman Process Model." IFAC-PapersOnLine 48.8 (2015): 309-314.

Debar, Hervé, Marc Dacier, and Andreas Wespi. "Towards a taxonomy of intrusion-detection systems." *Computer Networks* 31.8 (1999): 805-822.

J. J. Downs and E. F. Vogel. A plant-wide industrial process control problem. Computers & Chemical Engineering, 17(3):245–255, 1993.

Garitano, Iñaki, Roberto Uribeetxeberria, and Urko Zurutuza. "A review of SCADA anomaly detection systems." *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011.* Springer Berlin Heidelberg, 2011.

Hadžiosmanović, Dina, et al. "Through the eye of the PLC: semantic security monitoring for industrial processes." *Proceedings of the 30th Annual Computer Security Applications Conference*. ACM, 2014.

Kiss, Istvan, et al. "Data clustering-based anomaly detection in industrial control systems*." Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on.* IEEE, 2014.

Nardella, Davide "Snap7 package" http://snap7.sourceforge.net, 2015

Valli, Craig. "Snort IDS for SCADA Networks." (2009).

Verba, Jared, and Michael Milvich. "Idaho national laboratory supervisory control and data acquisition intrusion detection system (SCADA IDS*)." Technologies for Homeland Security, 2008 IEEE Conference on. IEEE*, 2008.

Yang, Dayu, Alexander Usynin, and J. Wesley Hines. "Anomaly-based intrusion detection for SCADA systems." *5th intl. topical meeting on nuclear plant instrumentation, control and human machine interface technologies (npic&hmit 05).* 2006.

Zhu, Bonnie, and Shankar Sastry. "SCADA-specific intrusion detection/prevention systems: a survey and taxonomy." *Proc. of the 1st Workshop on Secure Control Systems (SCS)*. 2010.

Junk Information in Hybrid Warfare: The Rhizomatic Speed of Social Media in the Spamosphere

Aki-Mauri Huhtinen and Jari Rantapelkonen
Finnish National Defence University, Helsinki, Finland
aki.huhtinen@mil.fi
jari.rantapelkonen@mil.fi

Abstract: The near exponential growth in social media (SM) communications is widely reported and services such as Twitter and Facebook duly have a combined user base in the billions. The growth of SM as a communication tool in recent years has also forced governments to consider the cyberspace as an important arena for strategic communication and disseminating their message. Further to this, if disseminated messages are misleading, distorted or false, a speedy response is required to limit the damage they can wreak. Recent events in Europe, such as the war in Ukraine and the annexation of Crimea, the Malaysia Airlines flight MH17 tragedy and the War in Syria, as revealed in reports by *The Interpreter* magazine, the Bellingcat open source investigation group and *Russia Today* (RT) respectively, show that misinformation is rife in SM. In this paper we analyze two case studies, namely Finland's Rapid Reaction Force and the Arrest of a Russian Citizen in Finland at US Request. We adopt a so-called rhizomatic focus to assess social networking spam and the consequences that this phenomenon creates for interaction in the security cases (Deleuze and Guattari 1983). In both case studies we analyze the respective timeline of events and the social media impacts on the rhizomatic "spam" information context. We argue that the rhizomatic way in which junk information spreads within social media is comparable to a 'spam world'. This spam world results from a technologized infrastructure that facilitates social media interaction without a proper understanding of the context of events.

Keywords: hybrid, rhizome, social media, spam

1. Nobody falls for spam or believes in trolls, do they?

SM has become a key player in hybrid warfare. (see Huhtinen and Rantapelkonen 2008) To this end, the authorities are finding it increasingly difficult to make a clear assessment of threats and to develop situational awareness on account of the rhizomatic character of information networks. The rhizomatic nature of the internet also confounds and confuses the ability to form a clear picture of events for the purposes of decision-making.

> *"Russian Twitter political protests ´swamped by spam´" (BBC 2012).*
>
> *"Hillary gets spam. Sorry… I mean, Russia attacks!!" (Poulsen 2015).*
>
> *"Hashtag Hijacked: Russia Trolls U.S. Twitter Campaign In Ukraine Crisis" (Johnson 2016)*

As the above quotes aptly illustrate, spam has evolved from junk email into junk messaging through social media. In this article, we use the term "spam" in the meaning of a junk or fake message targeted at readers, tweeters, and mass media consumers with the aim of arousing emotions, changing attitudes, and ultimately provoking a response. (see Chapple 2011)

Unsolicited email, or spam, used to be associated with automated posting in its early days. As almost everyone has been on the receiving end of automated spam at one time or another, one may ask whether anyone actually falls for it these days. Unfortunately, even those who are seasoned to receiving junk email can be susceptible. Spam continues to flood the net simply because it is still effective.

The term 'spam' is said to originate from the canned, processed 'fake meat' of the same name, immortalized in a famous 1970 Monty Python sketch set in a cafeteria, where the unwanted spam invariably pops up all over the menu and is chanted repeatedly by a group of Vikings who have invaded the café, drowning out the rest of the conversation (Hiskey 2010).

It has been argued that the first unsolicited messages came over the wires as early as 1864, when telegraph lines were used to send dubious investment offers to Americans. The first modern spam was sent on ARPANET, the military computer network that formed the technical foundation of the Internet. In 1978, a man named Gary Turk sent an unsolicited email to 400 people, advertising his new line of computers. In the 1990s, and specifically with the invention of social media platforms and services, a veritable Pandora's Box was opened. More recently, the nature of spamming has changed along with the influence it exerts (Fletcher 2009).

A new feature is the often-discussed issue of dramatic changes in communication technologies and the convergence of mass media with SM. According to Rheingold (2003), the ongoing communicational and high-tech wave "is the result of super-efficient mobile communications-cellular phones, wireless-paging, and Internet-access devices that will allow us to connect with anyone, anytime, anywhere". However, the real impact of this new technology depends on "how people use it, resist it and adapt to it".

In discussions about trolling, it has been suggested that humans are behind "social media spam", and not machines as such. A case in point was 55 Savushkina Street in St Petersburg, the headquarters of Russia's so-called "troll army", where bloggers were paid "to skew the truth and flood the internet with political innuendo" (Parfitt 2015), praising President Putin and denigrating the West.

In effect, the war in Ukraine has shown that social media spam such as junk tweets are part of the rhizomatic underground "war" being waged by the Twitter spammers. A problem arises when the mass media exposes these "underground wars" and creates hybrid rhizomatic wars. The rapid rate of posts on Twitter begs the question of whether the tweets in question were delivered automatically rather than by individuals. Russian Researcher Maxim Goncharov argues that "Whether the attack was supported officially or not is not relevant, but we can now see how social media has become the battlefield of a new war". To this end, Twitter accounts had been used to drown out chat and pollute the news stream (BBC 2012).

2. The Russian cybersphere of influence

The Kremlin's political goal is not expansion as such, but the preservation of sovereign autonomy in the face of the expansionist West (Morozov 2015, 27). Four out of every five Russian citizens are ethnic Russians, and this very ethnicity constitutes a massive political force. The remaining one-fifth is composed of over 150 different ethnic groups. Everyone in Russia has relatives or friends within their rhizomatic connections who are non-Russians, which poses a problem for the Kremlin: How to tap into the power of Russian-ness without damaging relations with non-Russian minorities at the same time? One option is for Russia to paint itself as a victim of Western policies and values by claiming that the West underestimates Russian culture and discriminates against Russian people. A second option is to emphasize cross-border solidarity and 'fellow countryman politics', or Pan-Slavism as it used to be known. A third possibility is for the Kremlin to emphasize the historical mission of Russians as the chosen people, thereby putting an end to Pan-Slavism.

Russia's isolationist policies and the increasing alienation of civil society against the Kremlin are also reflected in Russian interactions on the Internet, and there is a clear difference between the way citizens comport themselves in public and in private. The Internet, and particularly VKontakte, Russia's Facebook, have become key channels for self-expression in Russia. In 2014, in an attempt to tighten the government's already strong hold over the Internet, President Putin officially passed the so-called 'Bloggers Law', requiring any blogger with more than 3,000 readers to register with *Roskomnadzor*, Russia's media oversight agency. In practice, the Russian Armed Forces are also active in the information environment as they carry out tasks to contain and prevent military conflicts via the information environment (Russia 2011), which implies that this sphere is being militarized with the aim of controlling it.

The following example is illustrative of the Kremlin's trolling activities: "Since spring 2014, thousands of fake LiveJournal blogs have been mass-posting content promoting a pro-Kremlin stance on world events, attacking Western leaders and praising Russian president Vladimir Putin. Using custom Python code, [social sciences researcher] Lawrence Alexander was able to isolate and analyze these accounts. Delving deeper into the metadata of the supporting Twitter bot network could provide further clues as to their origin" (Alexander 2015). Another concrete example of rhizomatic networks at work is the Kaspersky Lab, a Moscow-based company currently ranking fifth in revenue among security software-makers worldwide. Founder and CEO Eugene Kaspersky was educated at a KGB-sponsored cryptography institute and went on to work for Russian military intelligence. The company now publishes reports on electronic espionage by the US, Israel, the UK and Russia (Matlack, Riley, & Robertson 2015).

The following sections focus on two further examples in the form of Finnish case studies, which clearly illustrate the way in which social media impacts the rhizomatic "spam" information context.

3. Case: Rapid reaction force

Finland's biggest daily newspaper, *Helsingin Sanomat*, reported on 25 June 2015 what would appear to be a routine story about a number of reservists in the Finnish Army being offered the chance to volunteer to be called up to active service faster than the law currently stipulates so as to create a rapid reaction force. The Finnish Ministry of Defence cited the increased security threat in Europe since the annexation of Crimea and the war in Eastern Ukraine as motivators for the decision. Later reports quoted the Finnish Minster of Defence, Jussi Niinistö, as saying both the Finnish Navy and Air Force already have this stipulation in place and the Army was following suit. (Defense News 2015).

The story attracted little attention outside of Finland for almost three weeks until 18 July when a US-based news service called "Defense News" ran the headline "Finland to deploy Quick Response Units Along Russian Border". The next day, the Russian News Agency "Novostimira" ran the story under the headline "Finland will deploy a rapid reaction force along border" (Novostimira 2015), quickly followed by several other Russian news outlets reporting on the same issue (Figure 1, Pillar 1).

Just under 24 hours later, a small number of Finnish press agencies reported on the claims circulating in the Russian media and in one instance corrected them, quoting a Finnish Ministry of Defence spokesperson (Figure 1, Pillar 2). At the same time, the Finnish media also issued stories quoting the Finnish Defence Minister, denying the claims in the Russian media (Figure 1, Pillar 3).

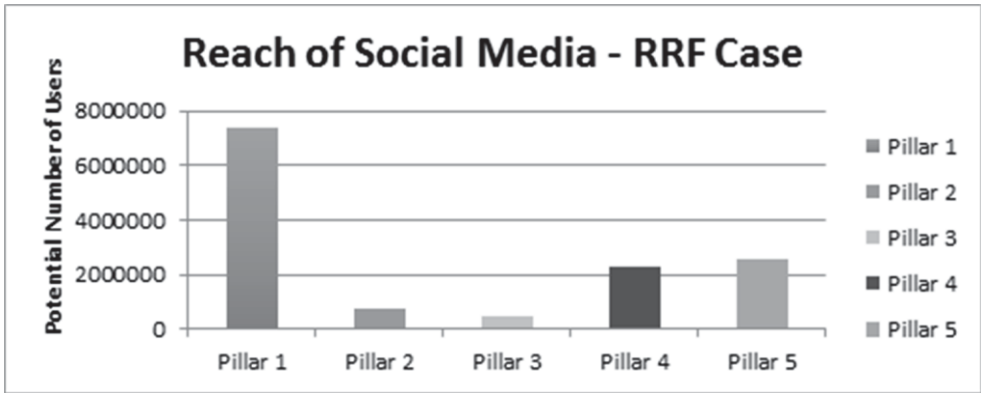


Figure 1: Reach of social media

On the evening of 21 July several Russian news agencies reported the Finnish correction (Figure 1, Pillar 4), while at the same time other Russian sites reported stories based on the original erroneous quote from Defense News (Figure 1, Pillar 5). For example, SputnikNews (2015) ran the headline "Scared Scandinavians: Finland to Militarize its Entire Border with Russia", with other unofficial blogs producing similar headlines. It seems reasonable to assume that such headlines represent a clear distortion of the facts. At first glance, this would seem to be a fairly straightforward case of a government's statement being reported by some media, misquoted by others, and then a correction being issued. However, by analyzing the reach of the social media coverage of the inaccurate information and the reach of the attempts to correct this, the report will show that the latter came a distant second in terms of reaching an audience. This indicates that attempts by Finland's Ministry of Defense to portray its actions as not being aimed at Russia largely failed in this case[1].

Let us now consider the timeline of events as commented upon on different social media platforms.

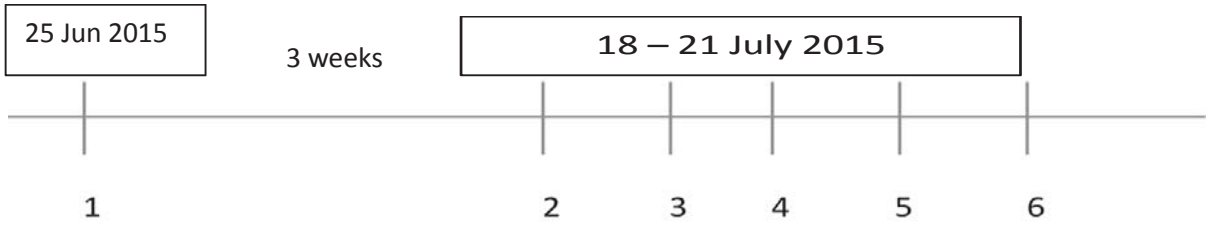


Figure 2: Finnish and Russian news outlets reporting timeline

[1] For discussion regarding reach of the social media posts, please see Research Methods section below.

- 1. 25 June 2015: Story breaks in Finnish media.
- 2. 18 July 2015: Article is reported by Defense News as militarization of the border.
- 3. 19 – 20 July 2015: Russian news quotes Defense News article.
- 4. 20 July 2015: Finnish Ministry of Defence issues statement correcting Defense News article.
- 5. 21 July: Certain Russian media quote Finnish correction under misleading headlines.
- 6. 21 July: Certain Russian media continue inaccurate reporting.

Key points to note from a social media perspective for this sequence of events are:

- The Finnish Defence Ministry posted two tweets related to this story, both as sub-tweets, adding to a total of 58 retweets 15 hours after Russian news ran their story[2].
- The Finnish Defence Ministry did not post a stand-alone tweet about this at any stage.
- The Finnish Defence Ministry spokesperson did not tweet about the issue at all.
- The Finnish Foreign Ministry did not tweet about the issue at all.
- The Finnish defense Minister tweeted from his personal account twice, 35 hours after the story broke at 1620 & 1959 Finnish time and received 21 retweets[3]
- The potential social media reach of social media users reading the inaccurate story in Russian social media was in the millions[4].
- As shown, it can be said with some confidence that the social media reach of the Finnish media reports was around 470,000 people[5].
- As shown in this case, the Finnish government's social media reach on this subject was in the low tens of thousands, significantly lower than the Russian counterpart and thus dwarfed by it.

This article has not considered the social media impact of the initial coverage of the story in June 2015 but has focused instead on the period from 19 – 21 July when the coverage quoting the inaccurate version of the story occurred. Thus, when referring to the story in the following sections, reference is made to the inaccurate claim that Finland was to deploy rapid reaction soldiers, not to the fact that Finland was enhancing capabilities.

4. Quantifying the social media impact

This story was run by 10 different Russian media outlets with the story being retweeted 211 times to a potential audience total of just over 2 Million followers. On Facebook and Vkontakte the potential reach was over 5.3 Million giving a total of around 7.7 million users. This can be seen at pillar 1 of table one below.

Against this, Finnish media ran 3 articles correcting the perception made in Russian news and then 3 articles about the Defense Minister's comments in which he explained that the Army was developing a capability and not deploying it. Taken together these articles reached of 255,000 on Twitter users, 200,000 on Facebook and slightly more than 17,000 people on Russian VKontakte. These were done 19 hours and 22 minutes after the first Russian articles appeared.

Once the Finnish Defense Ministry had issued a correction picked up by Russian media, 2 articles with the corrected story mentioned were run. In some cases these lead with headlines still indicating the deployment of troops, rather than an enhancement of capability (Figure 1, Pillar 4). While it cannot be said with certainty why, these appeared late in the evening away from prime time.
Subsequent to the corrected stories appearing, Kremlin funded media in at least one instance continued with inaccurate headlines, such as that by Sputnik mentioned above. This was significant as it had a potential reach of over 2.5 Million (Figure 1, Pillar 5).

---

[2] July 20th 10:19 [online] https://twitter.com/Puolustusvoimat/status/623029367019556864 & also July 20th 12:44: https://twitter.com/Puolustusvoimat/status/623065893749309441

[3] Tweets by Finnish Defense Minister [online] https://twitter.com/jniinisto/status/623537901406699521 & https://twitter.com/jniinisto/status/623482576968265728

[4] Russian News Social Media accounts (name & followers) RIA Novosti: 4 580 000, Novostimira: N/A, DP: 50 884, Fontanka: 93 346, Ruposter: 97 544, Sputnik: 2 585 700, NewsRu: 69 157, War and Peace: N/A, VZGLYAD: 95 285, Polit.ru: 100 295.

[5] Finnish agencies Yle & Helsingin Sanomat have 222 053 & 246 715 followers respectively.

It can be seen in the graph above that the efforts by the Finnish Ministry of Defense were much less effective on social media than the stories run by Russian News.

As illustrated in the graph above, the efforts by the Finnish Ministry of Defence were much less noticed on social media than the stories run by Russian news.

This case shows, as Mayfield (2011) also points out, that a fast response is an important aspect of social media communication. While the sheer size of the potential audience that the Russian media can reach may seem intimidating, the timings reveal that most of these stories were published over a period of around 6 hours (Figure 2), which seems to be a wide enough window to recognize the issue and issue a sustained, fast response using social media. The key, of course, lies in recognizing that it is happening, but that particular topic is outside the scope of this report.

Given that just two tweets were posted by the Finnish Ministry of Defence, with just one by the Minister of Defence, and both sometime after the story broke (in the case of the Minister, after 35 hours), it seems fair to argue that this was too few. In actual fact, waiting up to 35 hours is also too slow and effectively ceded the online space to those spreading, intentionally or otherwise, the inaccurate story. As a non-aligned country in the current security environment, it can easily be argued that its credibility was undermined by a slow and inadequate response in the online space.

The net result is that the numbers reached on social media in Pillar 4 quoting the Finnish government's response were rather low, especially when compared to the case which follows.

5. Case: Arrest of Russian citizen in Finland at US request

In the second case**,** on August 27 2015 it was reported in the Finnish media that the Ministry of Justice had, at the request of the United States and in accordance with a treaty between the two nations, arrested a Russian citizen in Helsinki wanted for crimes in the US state of Minnesota.

The case sparked a very swift reaction from the Russian Foreign Ministry, which sharply criticized the move and thus generated considerable online discussion on the subject. There would appear to be two broad alternatives that the reader could follow in viewing this case. Firstly, it can be said that Finland was operating in accordance with a transparent treaty and the accepted norms of international law or, alternatively, it can be viewed as yet more evidence of the US conducting what the Russian Foreign Ministry called a "witch hunt" (RT 2015). Clearly, it is in the interests of Finland for the former to become the main narrative and for this to be communicated well. Let us now consider how well this was achieved by examining the timeline of events.

Figure 3: Finnish and Russian news outlets reporting timeline

- 1. 26 Aug 2015 18:40: Russian Foreign Ministry issues statement condemning the arrest and calling it a witch hunt.
- 2. 26 Aug 2015 18:49: Both TASS and RT report statement at exactly the same time. Very wide coverage occurs in Russia.
- 3. 26 Aug 2015 20:54: Finnish news outlet *Ilta-Sanomat* reports the issue. Other Finnish outlets follow suit and report that an official Ministry of Justice statement will be made the following morning.
- 4. 27 Aug 2015 11:26. Finnish news reports Ministry of Justice statement confirming the arrest at the request of the US authorities in accordance with the treaty. Reuters later reports statement, which receives very wide circulation in social media.
- 5. 27 Aug 2015 11:33: Finnish Ministry of Justice tweets response in Swedish, Finnish, and English (at 12:13).

- 6. 27 Aug 2015 11:53: Russian news reports Finnish Ministry of Justice's statement and also appears to shift towards stating that the arrest was made at US request in accordance with the treaty.

As can be seen in the graph below, this story received wide attention in the Russian social media and, after the statement by the Ministry of Justice, achieved wide international reach as well.

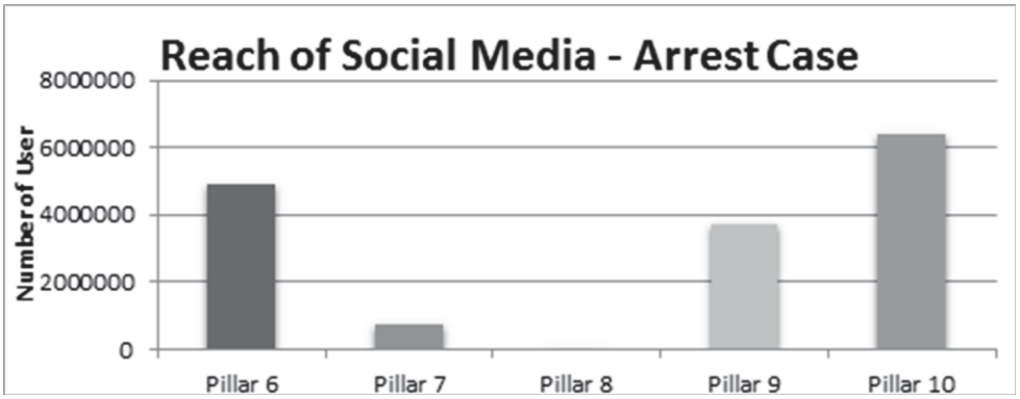


Figure 4: Reach of social media – arrest case

Figure 4 Pillar 10 indicates the swift official Finnish narrative as relayed by the Russian media. Pillar 9 shows the Finnish official narrative as relayed by the international media, excluding the Russian media. Pillars 6 and 7 (6 showing the response in the Russian media to the case and 7 the Finnish response) vary significantly in size due to the population differences between Finland and Russia (143 million vs. 5.4 million), but this does imply that 14% of Finns were reached as opposed to only 1.4% of Russians. Certainly, the fact that the Russian Foreign Ministry felt compelled to issue a statement could easily explain the interest in the Finnish media.

Pillar 8 is at such a low level as just 1,500 people follow the Finnish Ministry of Justice's Twitter account. The Finnish Foreign Ministry did not tweet or issue a statement as their Russian counterparts did, thus excluding their 50,000 followers from the case. The Russian Foreign Ministry Twitter accounts have over 800,000 followers.

The key point to observe is that on the evening of 26 August, the Ministry of Justice stated that an official response would be issued the following day. This was then made in English and allowed foreign news outlets to pick up on the story. A comparison of Pillars 9 and 10 with Pillars 4 and 5 in the previous example shows that swift official responses led to increased social media reporting of the narrative preferred by the Finnish government.

The fact that the international English language social media response to the Finnish Ministry of Justice's timely statement outweighed the Russian social media reaction to that same response is a key issue in this case. It seems fair to conclude that the swift and multilingual response by the Finnish government to the Russian government's statement largely explains this. It also seems fair to argue that by responding swiftly, it gave international news outlets the opportunity to present this information before their attention shifted to other events.

6. Research methods

The researchers adopted primarily a quantitative approach to gathering information on the reach of social data. When it came to understanding the nature and content of news certain articles, interpretative methods were more appropriate.

The quantitative methods involved using social media's advanced search function and other open source intelligence techniques. The technique was applied to each of the major Russian news outlets to identify which had made a social media post about their version of story (for example, in Case One, that the Finnish military was deploying troops rather than developing a capability) and noting the number of followers that each account had. Each post was then checked for the number of retweets or shares it received.

The same process was carried out for Finnish and other news agencies which tweeted versions of the story that reflected that government's position, corrections to articles in western media following the clarification issued by the Finnish side and then any corrections made in the Russian media.

An interpretive method was taken in determining the nature of the corrected versions issued by the Russian media so as to understand how well the remarks of the Finnish government had been incorporated into the story in each case. This was done to give further understanding of the impact of the corrected versions.

When gathering data about the reach of a story on social media, the researchers feel it is a reasonable assumption that an article shared by a social media account with a lot of followers (for instance several are in the millions) will be read more than one shared by one with fewer, such as the Finnish Defense Minister who has a around six thousand.

Further, without access to the analytics data for each of the media accounts in question, it is not possible to say for certain exactly how many of their followers viewed each post containing Russian versions of the story and how many viewed the Finnish versions. Therefore, a direct comparison of potential audiences was made to show the potential reach of each tweet. By doing this for two separate news events, a comparison can be made between the social media reaction to a delayed official response and to a more prompt one with all other factors being equal.

7. Discussion

By examining and comparing two case studies of news items reported in both the Finnish and Russian media, this article shows how social media exerts a huge impact on the mass media and vice versa. In practice, they have converged.

There is also clear evidence that in the absence of a swift official response to inaccurate news reports online, the speed at which social media can spread inaccurate information can lead to a misrepresentation of facts and intentions becoming the dominant narrative. This leads to the conclusion that institutions that wish to prevent their message from being undermined with such misinformation should be aware of how their message is being received in social media and react as quickly as possible to halt the spread of misinformation.

8. Conclusion

This article duly considered two separate cases occurring at similar periods in the Finnish news in 2015. In the first case, an enhancement of military capability was reported, first by a US-based news service and then extensively by Russia, as a deployment of military capability, which it clearly was not.

The slow official response to this inaccuracy ceded the social media space to reports that implied that Finland was preparing for imminent military action against Russia. It seems reasonable to argue that this was not a message that the Finnish government intended to send.

The second case, concerning the arrest of a Russian citizen in Finland at the request of the USA, provoked a much faster response. This allowed the Finnish government to claim a much larger share of the online space devoted to this story with their own narrative.

The second case clearly shows that a swift official response causes the social media space to be occupied more readily by the official narrative, in the sense that Finland was acting in a transparent, legal manner, rather than as a puppet of the USA.

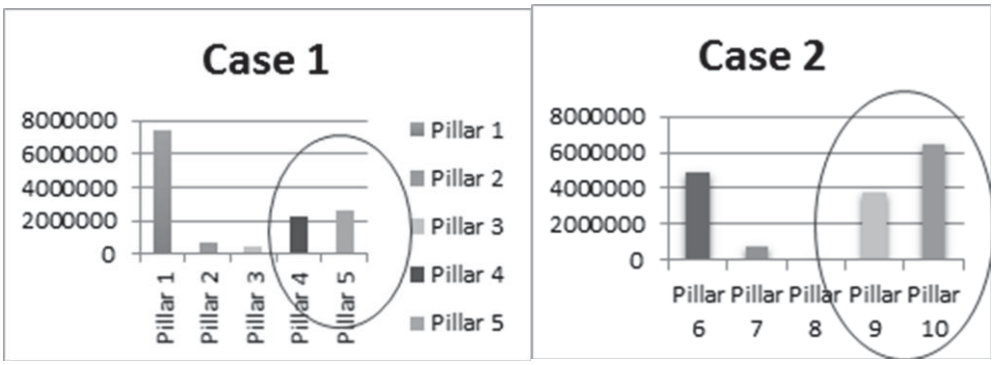


Figure 5: Both cases

Juxtaposing the two cases, as shown in Figure 5, reveals the potential reach of social media regarding each new development, indicating that the response to the Finnish government's position on social media, as highlighted, was reported much less in the first case (where the response was slower) than it was in the second case.

While further research across a larger number of cases is certainly called for, it seems reasonable to conclude from these cases that a swifter government response will very likely lead to media organizations presenting the information in question on social media more than they would in the face of a delayed official response.

In light of the findings, it can be concluded that spam, as defined as irrelevant or unsolicited messages sent over the Internet typically to large numbers of users for the purposes of propaganda and misinformation, obfuscates situational awareness among ordinary citizens and impedes government decision-making. By spamming multiple targets over a long period of time, spammers are able to gain currency for their messages and even inflict harm if they exert a sufficient influence over the target audience. The reasons for our susceptibility to distorted or false information are numerous, but one may be the individualistic nature of social media as a forum where all voices can be heard. Spam has been transformed from the admonishment "do not click on the links" to "click on the links and participate in the war of words". The more ambiguous or suspicious the news story, the stronger the pull to start trolling. Waging war has traditionally required time and patience but in a hybrid war fought in a world of spam those factors are no longer relevant. Knowledge is no longer power as junk information can sometimes gain the upper hand if matters are not dealt with swiftly and accurately on all informational fronts.

Acknowledgements

The authors acknowledge the valuable contributions of Mr. Ian Robertson and Mr. Joonas Vilenius of WG Consulting, a social media intelligence and research consultancy. They can be contacted at *WGConsulting@protonmail.com*

References

Alexander, L. (2015) Massive LiveJournal Troll Network Pushes Pro-Kremlin Narratives. Stopfake.org, Dec 24, 2015 [online], http://www.stopfake.org/en/massive-livejournal-troll-network-pushes-pro-kremlin-narratives/.

BBC (2012) 'Russian Twitter political protests "swamped by spam"', BBC, March 8, 2012, [online], http://www.bbc.com/news/technology-16108876.

Chapple, M. (2011) 'Anatomy of a Spam Attack'. *BizTech*, Dec 7, 2011, [online], http://www.biztechmagazine.com/article/2011/12/anatomy-spam-attack.

Defense News (2015) 'Finland To Deploy Quick Response Units Along Russian Border'. *Defense News*, July 18, 2015. [online], http://www.defensenews.com/story/defense/land/army/2015/07/18/finland-tactical-response-units-russia/30197931/.

Deleuze, G. and Guattari, F. (1983) *On the Line*. Translated by John Johnston, Semiotext(s), New York.

Fletcher, D. (2009) 'A Brief History of Spam'. *Time*, Nov 2, 2009, [online], http://content.time.com/time/business/article/0,8599,1933796,00.html.

Hiskey, D. (2010) 'How the Word "Spam" Came to Mean "Junk Message"' *Today I Found Out*, Sept 10, 2010, [online], http://www.todayifoundout.com/index.php/2010/09/how-the-word-spam-came-to-mean-junk-message/ .

HS (2015) 'HS: Finnish Defence Forces set up new rapid deployment force'. June 25, 2015, [online], http://yle.fi/uutiset/hs_finnish_defence_forces_set_up_new_rapid_deployment_force/8103168.

Huhtinen, A-M. and Rantapelkonen, J. (2008) *Messy wars*. Finn Lectura, Helsinki.

Johnson, L. (2014) 'Hashtag Hijacked: Russia Trolls U.S. Twitter Campaign In Ukraine Crisis'. *RFERL*, Jan 22, 2016, [online], http://www.rferl.org/content/ukraine-us-russia-twitter-trolling/25362157.html.

Matlack C., Riley, M. and Robertson, J. (2015) Kaspersky Lab has published reports on alleged electronic espionage by the U.S., Israel, and the U.K.— but hasn't looked as aggressively at Russia, [online], http://www.bloomberg.com/news/articles/2015-03-19/cybersecurity-kaspersky-has-close-ties-to-russian-spies.

Morozov, V. (2015) 'Aimed for the Better, Ended up with the Worst: Russia and International Order', *Journal on Baltic Security* Vol 1, Issue 1, 26–36.

Novostimira (2015) Finland will deploy a rapid reaction force along the border. Novostimira, July 19, 2015.

Parfitt, T. (2015) 'My life as a pro-Putin propagandist in Russia's secret "troll factory"'. *The Telegraph*, June 24, 2015, [online], http://www.telegraph.co.uk/news/worldnews/europe/russia/11656043/My-life-as-a-pro-Putin-propagandist-in-Russias-secret-troll-factory.html.

Poulsen, K. (2015), [online], https://twitter.com/kpoulsen/status/649379405714710530.

Python, M. (2007) SPAM: The origin of the word (Monty Python sketch), [online], https://www.youtube.com/watch?v=Ur5nE5uTa6s.

Rheingold, H. (2003) *Smart Mobs: The Next Social Revolution*. Basic Books, MA, USA.

RT (2015) '"Witchhunt": Moscow slams Finland's arrest of Russian citizen on US request'. *Russia Today,* Aug 27, 2015, [online], https://www.rt.com/news/313648-finland-russia-arrest-senakh/.

Russia (2011) *Conceptual Views on the Russian Armed Forces' Activities in the Information Environment*. The Ministry of Defence of the Russian Federation.

Sputnik News (2015) 'Scared Scandinavians: Finland to Militarize its Entire Border With Russia', *SputnikNews*, July 21, 2015, [online], http://sputniknews.com/europe/20150721/1024877311.html

Modeling the Impact of Cyber Risk for Major Dutch Organizations

Vivian Jacobs, Jeroen Bulters and Maarten van Wieren
Deloitte Cyber Risk Services, Amsterdam, The Netherlands
vjacobs@deloitte.nl
jbulters@deloitte.nl
mvanwieren@deloitte.nl

Abstract: We have developed a generic model to determine the Cyber Value at Risk for individual organizations that provides insight into the magnitude of cyber risk as well as the factors driving the risk. This model distinguishes four types of threat actors and seven types of information assets for each organization. A heuristic approach is adopted where gaps in the available data are filled in with estimations based on expert judgement. We have applied this model to a group of major Dutch organizations that jointly represent major sectors in the Dutch economy, including the Public Sector. This determines, for each of the considered sectors, the expected financial impact as well as the Cyber Value at Risk ("worst-case" scenario, or the loss that will annually not be exceeded with a likelihood of 95%). The yearly value loss for the entire Dutch economy is expected to be approximately €10bn. The cyber risk exposure is highest in the Banking, Defense and Aerospace, Technology and Electronics sector and the Public Sector.

Keywords: cyber risk, cyber attack, risk quantification, information security, value at risk, risk management

1. Introduction

Quantifying risks associated to cybercrime is a difficult task. The two main causes of this are, on the one hand, the scarcity of suitable data on cyber incidents, and on the other hand, the lack of a universal standardized framework to assess cyber risk. The notion of Cyber Value at Risk initiated by the World Economic Forum aims to provide such a framework (World Economic Forum, 2015). Further developed by Deloitte, the Cyber Value at Risk (VaR) model is a first step in understanding the origin and magnitude of cyber risk for individual organizations on a stand-alone basis. Existing quantification methods often focus on the technical aspect of cyber risk (Dudorov, Stupples and Newby, 2013; Raugas et al., 2013). The Cyber VaR model unifies technical aspects of cyber security with the business and management considerations of an organization. This allows organizations to systematically understand their exposure to cyber risk and make sound decision on investments regarding cyber defense.

As a feasibility demonstration, the Cyber VaR model is applied to the largest organizations for the fourteen sectors in the Netherlands that give rise to the largest amount of cyber risk. These organizations (about 50 in total) also represent the largest part of the Dutch economy. In particular, the Cyber VaR model determines the expected financial value loss of a cyber incident, as well as the limit to financial value loss in a 95% confidence interval (or Cyber VaR) for the 14 sectors considered. Furthermore, it gives insight into the origin of cyber risk by specifying which information assets of a sector are targeted by which threat profiles.

Lack of data is a fundamental problem since cyber incidents are known to be underreported (Verizon, 2015), and the incidents that get reported are usually not accompanied by full details. However, there is a lot we do know about cyber incidents on a qualitative basis. Our approach is to translate this qualitative understanding into a generic and simple model that captures the logical structure of interaction between an organization and its cyber-attacker(s). The design of the model is such that the few parameters that are unavoidably introduced can be easily understood, interpreted, and reasoned about.

These parameters are subsequently estimated on the basis of relevant data (Statistics Netherlands, 2014; Elsevier, 2015; Verizon, 2015; Statistics Netherlands et al., 2015; CGI, 2015) as well as expert judgement where no concrete information is available (Tetlock and Gardner, 2015). To this end, Deloitte security professionals with extensive experience in cyber risk management, as well as academic and government experts in cyber security have been involved in validating the assumptions underlying these estimates. Furthermore, both observed fluctuations and the uncertainty caused by not precisely knowing parameter values contribute to the Cyber Value at Risk.

In this article we focus on the methodology of the Cyber VaR model, which will be explained in the section Approach and Methodology. After this, we present our main findings in the section Results.[1] We end with a Discussion.

2. Approach and methodology

In explaining our approach and methodology in more detail, we explain first the high level model and subsequently each of its components. Finally, we discuss the selection of sectors for our analysis.

2.1 Cyber value at risk model

The Cyber Value at Risk model consists of three core components and their interactions, illustrated in Figure 1. First of all, the attacks and the threat actors carrying them out are classified into four distinct *threat profiles*. Next, the attack mitigation capabilities for each individual organization (i.e. its *cyber security* maturity) are determined. Finally, seven types of *information assets* are defined. The information assets represent different assets of value for the organizations in each sector. Information assets are subject to various types of cyber-attacks. The maturity of each sector's cyber security determines how well each type of information asset is defended against these attacks. Furthermore the financial impact of a given attack on a certain information asset is determined. Finally, two uncertainty scenarios are considered: the expected-case and the worst-case scenario.

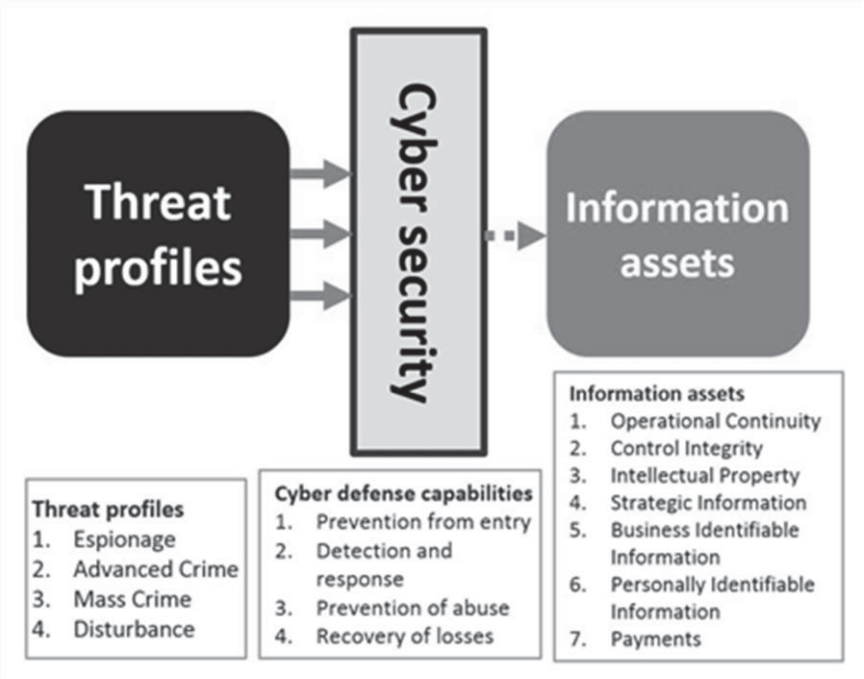


Figure 1: The three core components of the Cyber VaR model are the four threat profiles, the four cyber defense capabilities and the seven information assets listed here

The interaction between the Cyber VaR model components can be summarized as follows.

- The value of each information asset is determined per sector by considering what happens in case that this information asset is abused in its entirety. Given that we only make use of public data, it is assumed that this impact scales with an organization's financial parameters (i.e. reported data) and sector-specific properties. In addition, we consider the costs of fines and claims in case of abuse.
- Threat profiles are attracted to these information assets. The attractivity of information assets to a threat profile is proportional to the information assets' value and determines how the totality of threat actors spreads out over the information assets.
- The incentive for cyber spies may differ from a purely financial one. Therefore the Espionage profile is assigned additional attractivity to the Strategic and Intellectual Property information assets in specific sectors.

[1] The Cyber Value at Risk model was subject to an online publication by Deloitte on April 4, 2016 (Deloitte, 2016). In this report, more results for all the sectors can be found.

- When an organization's security system has been compromised, the impact of abuse accumulates until either the information asset is entirely abused, or the threat actor is detected and neutralized.
- At each stage of the calculation we keep track of the average impact and the Cyber Value at Risk, i.e. the impact in the most severe scenario.

These statements are explained in more detail now.

2.2 Uncertainty and Cyber VaR

Notwithstanding our careful considerations, making estimations means that a degree of uncertainty remains. Part of this is caused by a lack of data, which can be reduced in the future by obtaining additional data and stimulating high quality data collection methods. However, we had to deal not only with measurement uncertainty, but also with natural parameter fluctuations and unpredictability of human behavior. We have interpreted all these forms of uncertainty as contributing factors to the risk.

Whereas the average impact describes the expected situation, the Cyber VaR describes the most severe situation being the value loss that is yearly not exceeded with 95% probability. In other words, the Cyber VaR is expected to be exceeded only once in 20 years. For a single organization, knowing the average impact is not enough since the worst-case impact can significantly differ from the average impact. This difference in itself is an important aspect in determining a risk management strategy. The Cyber Value at Risk multiplier is the ratio of the value loss in the worst-case scenario and the expected value loss. The uncertainty of the estimated parameters is transformed into a bandwidth relating the expected situation and worst-case scenario. The propagation of these uncertainties is taken into account throughout the calculation.

The model is necessarily an a priori construction. Empirical evidence of the Cyber VaR results may be unavailable and is in many cases nonexistent, because part of the (risk leading to) value loss will be overlooked. This may be seen as a downside of the model, but disregards the function and rationale of Value at Risk. Our justification is that we have provided the most logical structure possible on well-defined objects (the model components) that can be mapped to the few available observations in order to calibrate parameters, all of this within an uncertainty that is translated into risk itself.

2.3 Information assets and value impact

Seven types of information assets, representing the major forms of value to an organization known to us that can be abused by cyber attackers are identified. These information assets and a description can be found in Table 1.

Table 1: The seven information assets in the Cyber VaR model, their main value contributors and their threat description

Information assets	Main value driver and description of threat impact
Operational Continuity	Income: loss of daily income when ICT or operations systems are unavailable
Control Integrity	Assets: loss of control over non-cash assets or products
Intellectual Property	Equity: loss of competitive advantage from investments in IP
Strategic Information	Growth: loss of M&A opportunities after leak of confidential data
Business Identifiable Information	Market share: loss of clients after leak of confidential third-party information
Personally Identifiable Information	Market share: loss of customers and employees after leak of confidential data
Payments	Liquidity: cash loss from fraudulent execution or modification of financial transactions

These information assets can be assigned a value, based on the organization's revenue, liquidity and equity. The components driving the value of each information asset are listed in Table 1. Thus, the actual value is computed differently for each information asset. This is the value impact for an organization when the asset abused in its entirety. This impact may be noticed immediately, through loss of cash or daily income when Payments or Operational Continuity are abused, respectively. In other cases, it may take some time before the value impact

materializes. This is for example the case for stolen Intellectual Property whose future value is lost regardless of this delay in detection. Finally, in some cases the impact value is not realized at all, when attacks remain undetected or unreported. Besides the base value of information assets, additional value impact comes from associated claims and fines, for instance when confidential information is published.

2.4 Threat profiles

The cyber attacks are distributed over the threat profiles and information assets in the following way. Cyber attackers typically exhibit two kinds of behavior: a slow and a fast approach. This bifurcation between slow and fast attackers has been reported in the literature and was also recently modeled (Van Wieren et al., 2016). The logic behind this is the fact that attackers can optimize their gain in two ways: either by acting fast and unconcealed, and grabbing all the data they can get before being neutralized by the defense system, or by infiltrating and abusing the system at a very slow rate, staying under the radar for as long as possible. Following these observations two attacker sophistication levels are identified. Fast attacks are assigned a low sophistication, while a slow approach is associated with a high sophistication. Thus, low sophistication attackers will be detected in a timely manner and must strive to maximize their gains by acting as fast as possible. Highly sophisticated attackers have ways to remain undetected for a longer period and can spend larger amounts of time (and money) to target very specific information assets or to prevent detection.

Based on reports about the (Dutch) cyber threat landscape (AIVD, 2015; ENISA, 2015), attackers of both sophistications are also classified by their incentives, which results in the four threat profiles mentioned in the beginning of this section. These four profiles are: Espionage, Advanced Crime, Mass Crime and Disturbance. They are described below, a summary can be found in Table 2.

Table 2: List of threat profiles used in the Cyber VaR model, their sophistication and activity level, and the main information assets targeted by each profile

Threat profiles	Main information assets targeted	Sophistication	Activity
Espionage	Strategic Information, Intellectual Property	High	Slow
Advanced Crime	Payments, Strategic Information		
Mass Crime	Payments, Personally Identifiable Information	Low	Fast
Disturbance	Operational Continuity		

2.4.1 Description of threat profiles

Attackers in the Espionage and Advanced Crime profiles need to act slowly to remain stealthy. Both these profiles include organized groups who prepare their big attacks well and use highly sophisticated methods. Espionage includes corporate espionage but also hacker teams associated to nation states, and is mainly attracted to targets of strategic value, i.e. Strategic Information and Intellectual Property. In contrast, Advanced Crime attackers seek primarily financial gain, targeting the Payments information asset and, in case of insiders, Strategic Information.

The Mass Crime and Disturbance profiles have a low sophistication, their high abuse rate makes it unnecessary to stay undetected. The Mass Crime profile is the most prevalent. These attackers do not target specific assets or organizations but have a widely deployable toolbox, taking advantage of known vulnerabilities and delays in the defenders' response. Examples are usage of phishing, malware, and ransomware (ENISA, 2015). While these attackers preferably target the Payments and Personally Identifiable Information assets for their own financial gain, a consequence of Mass Crime attacks may be system overload leading to compromised Operational Continuity and Control Integrity.

Finally, the Disturbance profile contains hackers with ideological or political motivations. Their goal is disruption of an organization's Operational Continuity. To this end, attackers in the Disturbance profile try to compromise the integrity or availability of critical systems, or publish privacy-sensitive data.

The total number of attacks is divided over these 4 threat profiles according to Table 3. The attackers belonging to a certain threat profile are further distributed over the information assets according to their respective attractivity.

Table 3: Distribution of attacks over the threat profiles. We have also added the best-case scenario. These numbers are assumptions based on our research and are used as input for our model

Number of attackers/year	Espionage	Advanced Crime	Mass Crime	Disturbance
Best-case number	650	100	16000	125
Expected number	1000	200	20000	500
Worst-case number	1500	400	24000	2000

2.4.2 Attractivity of information assets

The attractivity of information assets to each of the threat profiles, mentioned earlier, is shown in Table 4. Information assets can have a low (L), medium (M) and high attractivity (H), respectively. Going from L to M and from M to H makes the information asset a factor 7 more attractive each time. Aside from this "base-level" attractivity, some information assets in specific sectors have additional attractivity to the Espionage profile due to their strategic value (not shown in Table 4). The combination of the attractivity matrix in Table 4, the distribution of threat profiles in Table 3, and the sophistication levels in Table 2 fixes the relation between the "Threat profiles" core component and the other two components of the model.

Table 4: Matrix of attractivity of the information assets to the threat profiles

Attractivity matrix	Espionage	Advanced Crime	Mass Crime	Disturbance
Operational Continuity	L	L	M	H
Control Integrity	M	L	M	M
Intellectual Property	H	L	L	L
Strategic Information	H	M	L	L
Business Identifiable Information	M	L	L	L
Personally Identifiable Information	M	L	M	M
Payments	L	H	H	L

2.5 Cyber defense capabilities

Each organization has a specific defense system protecting their information assets, which is modeled by the four cyber defense capabilities summarized in Table 5. All sectors have a base defense level for each of these capabilities. This is the estimated maturity based e.g. on the prehistory this sector has with cyber attacks.

A single attack is then modeled as a 3-step process as shown in Figure 2. At every stage of the attack, information about the organization's cyber defense capabilities is combined with the sophistication of the attacker. Here it is assumed that for a highly sophisticated attacker, the organization's defense system is less difficult to compromise than for an attacker of low sophistication. The Non-criminal assessment stage consists of all hackers who have not (yet) gained entry to an organization's systems, and is associated to the Prevention from entry capability. If an attacker has a high enough sophistication, it can enter the second stage, Criminal Assessment, where the capabilities Detection and response, and Prevention of abuse are involved. The sophistication of the attacker versus the Detection and response time then determines how long attackers stay in the final stage and are able to abuse information assets. In this stage, the value loss starts accumulating in a way proportional to the number of days the attacker is able to abuse information assets, and the abuse rate (related to the activity, see Table 4). Finally, after the attacker is satisfied or neutralized, the accumulated loss can be reduced somewhat by the Recovery of losses capability. We have also taken into account a possible extra fortification of particular information assets by the organization. For example, organizations in the financial sector have (without notable exceptions) put additional protections in place for the Payments information asset.

Table 5: List of defense capabilities introduced in the Cyber VaR model and their description.

Cyber defense capability	Definition
Prevention from entry	Fraction of attackers blocked from entering an organization's system
Detection and response	Number of days before the attack is discovered and the threat neutralized
Prevention of abuse	Number of days between a hacker gaining access to the system and actually abusing information assets
Recovery of losses	Fraction of loss recovered by damage reduction capabilities of an organization

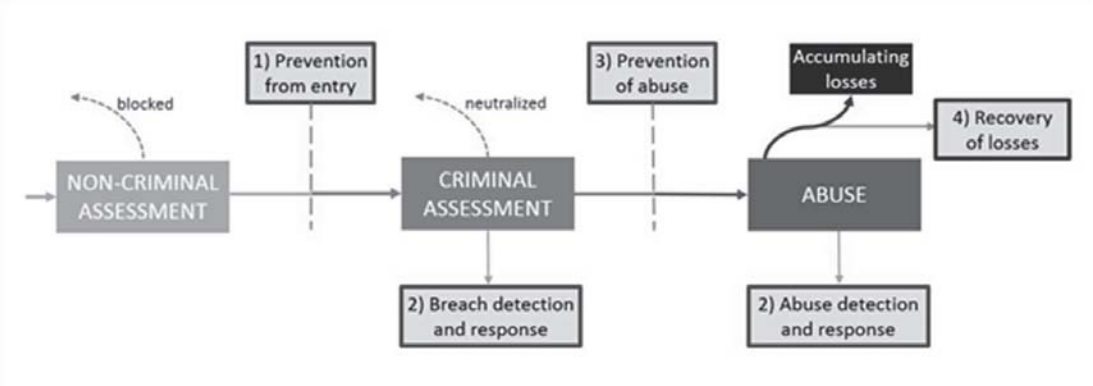


Figure 2: Attack-defense process in the Cyber VaR model, having three stages and involving the organization's four cyber defense capabilities

2.6 Dutch sectors and organizations

2.6.1 Description of sectors

The analysis includes the 14 largest sectors in the Netherlands in terms of income and risk level, listed in Table 6. The gross income is determined from annual statements of the largest organizations within each sector. Smaller sectors are not included. Organizations in Healthcare, Education and Central Government together comprise the Public Sector. These three components have been analyzed separately but results are presented under the Public Sector heading for conciseness. Public Sector does not include the Defense component. An estimation of the government's Defense expenditure is instead included in the Defense and Aerospace sector. The Insurance, Banking, and Asset Management and Pensions sectors are usually merged to form the Financial Services sector. We have chosen to keep these three components as separate sectors. The reason is their differing business models, which also lead to different risk profiles.

2.6.2 Sector-wise cyber defense maturity

Determination of the maturity of Dutch organizations' cyber defense is not straightforward. We have inferred the average maturity of each sector from literature data and validated this with experts in the field. This results in a typical maturity range for each sector as can be seen in Table 6. In general cyber maturity is measured on a 1 to 5 scale (1 representing non-existent, 5 representing a fully defined and optimized capability). Individual organizations within a sector may have a slightly higher or lower maturity than the sector average. Large organizations in terms of gross income have an on average higher maturity than small organizations. Relatively mature sectors are the Defense and Aerospace, Banking, and Oil, Gas and Chemicals sectors. The maturity of the Public Sector varies heavily over its components but is highest for Central Government. Relatively low maturity is assigned to organizations in the Education and Healthcare components of Public Sector and to the Utilities sector.

Table 6: Dutch sectors, their gross income and average maturity of the cyber defenses in these sectors. The gross income figures per sector are based on public data from Statistics Netherlands (Statistics Netherlands, 2014) and Elsevier (Elsevier, 2015). In an iterative process we have constructed a slightly adapted list of sectors to manifest the differences in these sectors' risk profiles. The cyber maturity range of each sector is estimated based on public reports and white papers (Verizon, 2015; Statistics Netherlands et al., 2015; CGI, 2015), as well as discussions with security experts. The figures on base value drivers in each sector are assumptions based on (financial) reports of the largest or most representative organizations in each sector, that are subsequently validated in interviews with professionals having experience in this sector

Sector	Gross income (€bn)	Cyber maturity range	Base value impact drivers		
			Market share and IP (income)	Controls and products (assets)	Strategic (equity)
Oil, Gas and Chemicals	720	3 – 5	0%-13%	0%-0.04%	2%-13%
Public Sector	389	1 – 4	0%-10%	0%-0.2%	0%-3%

Sector	Gross income (€bn)	Cyber maturity range	Base value impact drivers		
			Market share and IP (income)	Controls and products (assets)	Strategic (equity)
Wholesale and Retail	245	2 – 4	0%-5%	0%-1%	0.3%-2%
Asset Management and Pensions	227	1 – 4	0%-3%	0%	0%-0.3%
Insurance	141	2 – 4	0%-13%	0%	0.3%-2%
Consumer Goods	130	2 – 4	0%-10%	2%-6%	1%-8%
Banking	99	3 – 5	0%-50%	0%	1%-5%
Telecom	65	3 – 5	2%-50%	10%-40%	1%-4%
Technology and Electronics	42	2 – 4	0%-38%	2%-6%	1%-4%
Business and Professional Services	36	2 – 4	2%-100%	0%	0%
Transportation	33	2 – 4	0%-5%	0%-1%	0%-0.3%
Media	27	2 – 4	0%-5%	2%-6%	1%-4%
Utilities	25	1 – 3	0%-5%	0%-1%	0.3%-2%
Defense and Aerospace	20	3 – 5	0%-100%	25%-100%	40%-100%

2.6.3 Sector-wise base value impact factors

The differences between sectors in terms of value loss from cyber abuse can be small, such as for operational continuity or payments. For other types of abuse, there may be significant variations between the sectors. In Table 6, the most important differences are displayed for impact per value impact driver. These value impact drivers can be connected to the information assets by looking at the information asset descriptions in Table 1. This concerns impact on market share (through abuse of Business Identifiable Information or Personally Identifiable Information or IP), impact on asset value (through abuse of Control Integrity on assets or products) and finally, equity value (through abuse of Strategic information). The value impact from cyber abuse will depend on the details and these have been translated into uncertainty bands for the relevant parameters. In determining these factors, we have taken into account the business model, the nature of the information asset as well as the behavior of customers (market share), the intrinsic product value (controls and products) and the volatility, growth and M&A activity for the relevant sector (equity). These factors are, within generally accepted accounting principles, considered to be the major contributors to an organization's valuation.

3. Results

The main results of the analysis are the expected value loss and Cyber Value at Risk for the 14 sectors. Table 7 shows these two quantities. The columns labeled "absolute" give the value loss in billion €. The relative impact for each sector is the absolute value loss divided by the total revenue of this sector. The ratio between the expected absolute value loss and the absolute Cyber VaR is the Cyber VaR multiplier, in the rightmost column. The relative Cyber VaR in the third column of Table 7 can be attributed to each information asset, the result of which is shown in Table 8.

Table 7: Results for expected impact and Cyber VaR for all sectors, all values are rescaled to the size of the entire Dutch sectors

Value loss per sector	Total expected value loss (€bn)	Expected impact as ‰ of gross income	Absolute Cyber VaR (€bn)	Cyber VaR as ‰ of gross income	Cyber VaR multiplier
Oil, Gas and Chemicals	2.4	3.3	22	30	9
Public Sector	2.4	6.1	25	63	10
Wholesale and Retail	1.4	5.5	6.5	27	5
Asset Management and Pensions	0.2	0.9	1.1	4.8	6
Insurance	0.3	2.3	3.4	24	11
Consumer Goods	0.4	3.1	0.6	4.8	2
Banking	0.4	3.6	6.5	66	18
Telecom	0.3	4.9	1.7	26	5
Technology and Electronics	1.1	26	7.1	169	7
Business and Professional Services	0.3	9.4	1.8	49	5

Value loss per sector	Total expected value loss (€bn)	Expected impact as ‰ of gross income	Absolute Cyber VaR (€bn)	Cyber VaR as ‰ of gross income	Cyber VaR multiplier
Transportation	0.2	4.6	0.8	23	5
Media	0.0	1.4	0.3	12	8
Utilities	0.2	6.8	1.2	48	7
Defense and Aerospace	0.4	21	3.3	167	8

Table 8: Relative Cyber VaR per sector and information asset.

Cyber VaR as ‰ of gross income	Operational Continuity	Control Integrity	Intellectual Property	Strategic Information	Business Identifiable Information	Personally Identifiable Information	Payments	**All information assets**
Oil, Gas and Chemicals	21	0.7	2.3	4.3	0.7	0.6	0.0	**30**
Public Sector	23	26	4.4	7.7	0.1	2.1	0.2	**63**
Wholesale and Retail	18	0.8	0.0	0.0	0.0	7.9	0.0	**27**
Asset Management and Pensions	1.1	0.9	0.0	0.0	0.4	2.6	0.0	**5**
Insurance	4.5	0.5	0.0	0.1	2.5	16.5	0.0	**24**
Consumer Goods	1.3	1.5	0.9	0.3	0.0	0.8	0.0	**5**
Banking	2.3	1.5	0.0	0.6	44	13	4.8	**66**
Telecom	9.4	11	2.1	0.1	0.8	3.2	0.0	**26**
Technology and Electronics	14	26	129	0.1	0.2	0.4	0.0	**169**
Business and Professional Services	12	0.3	2.1	0.0	31	3.9	0.0	**49**
Transportation	23	0.7	0.0	0.0	0.2	5.5	0.0	**23**
Media	2.6	7.4	0.7	0.4	0.2	0.4	0.0	**12**
Utilities	23	20	0.0	0.3	0.2	5.5	0.0	**48**
Defense and Aerospace	0.5	49	27	89	1.2	0.2	0.0	**167**

Furthermore, Table 9 shows the fraction of each information asset contributing to the total Cyber VaR of all organizations.

Table 9: Attribution of information assets to the Cyber VaR of all sectors, resulting from the analysis.

Information asset	Fraction
Operational Continuity	41%
Control Integrity	18%
Intellectual Property	12%
Strategic Information	10%
Business Identifiable Information	8%
Personally Identifiable Information	10%
Payments	1%

Finally, Table 10 shows the total attribution of the threat profiles to the Cyber VaR.

Table 10: Attribution of threat profiles to the Cyber VaR of all sectors, a result from the analysis.

Threat profile	Fraction
Espionage	47%
Advanced Crime	1%
Mass Crime	41%
Disturbance	11%

From these results the following conclusions can be drawn. The highest cyber risk exposure is found in the Public Sector, the Banking sector, the Defense and Aerospace sector, and the Technology and Electronics sector. For Defense and Aerospace, and Technology and Electronics which are smaller sectors in terms of total revenue, the high exposure comes from the high Cyber Value at Risk per revenue of these sectors. For the Banking sector, which has a relatively mature defense system, the expected value loss is small. As a consequence, the Cyber VaR multiplier is relatively large for the Banking sector. In loose terms, their base defense is so good that when the worst-case scenario takes place, it will be extra painful. Finally, the Public Sector has both a relatively high Cyber VaR and a high Cyber VaR multiplier.

The Cyber VaR multiplier is for most sectors of order 10. This means that the uncertainty associated to cyber risk exposure is on average not extremely large. A worst-case scenario does have a larger impact, but will in most cases not destroy an organization completely.

The expected impact for the Dutch economy as a whole is of the order of 10 €bn each year. This number is the sum of the expected value loss for every sector. We have not taken into account interdependencies between organizations but in reality these will affect the expected impact.

4. Discussion

This article presents a proof of principle in establishing Cyber VaR as a framework for quantification of cyber risk. The approach and study presented here give insight in the relation between the technical, management and economic implications of cyber risk. However, room for improvement of the Cyber VaR model presented here remains. Most importantly, the lack of data and associated assumptions have already been mentioned. Over time, we expect to acquire both more as well as better quality data, and thus further refine the parameter estimations, leading to uncertainty playing a smaller role in Cyber VaR in the near future.

Furthermore, we should stress that interdependencies between organizations have not been taken into account. In the analysis, organizations are considered as stand-alone. In reality, there will be correlations and diversification effects because organizations share the same infrastructure and information, thus the risk exposure will also be shared. In the future, we also intend to address this issue by structurally taking into account these second-order effects. Especially regarding critical infrastructure, this is expected to have significant impact.

The scope of our work includes only the largest organizations in the Netherlands. Smaller sectors or sectors with a negligible risk exposure have not been taken into account. Examples are the Construction, Real-estate, Local Government and Leisure sectors. Furthermore, we have only considered economic impact, whereas values other than financial may be of relevance, for example public safety and integrity of the legal system. It may be interesting to also consider these non-economic values. Finally, also natural disasters, human error and malicious insiders were not in the scope of this analysis.

References

AIVD (2015), “Cybersecuritybeeld Nederland 2015”, [online], https://www.aivd.nl/binaries/aivd_nl/documenten/kamerstukken/2015/10/14/cyber-security-beeld-nederland-2015/csbn-oktober-2015.pdf

CGI (2015), “Is a Cyber Breach Inevitable? Cyber Security Challenges in the Netherlands”, [online], http://automatie-pma.com/wp-content/uploads/2015/05/CGI-Cyber-Security-White-Paper-Final.pdf

Deloitte (2016), “Cyber Value at Risk in The Netherlands”, [online], https://secure.myclang.com/3/4/187/1/o-oL6FYQttuBmi3djL6tN5EYFV06gZk45rpEh3hl8qljlYvXkeeYir9rNWkBvk5

Dudorov, D., Stupples, D. and Newby, M. (2013) “Probability Analysis of Cyber Attack Paths against Business and Commercial Enterprise Systems”, *Conference Proceeedings*, 2013 European Intelligence and Security Informatics Conference, Uppsala, Sweden, 12 – 14 August 2013, pp. 38 – 44

Elsevier (2015), “Top 500 grootste bedrijven”, [online], http://onderzoek.elsevier.nl/onderzoek/top-500-bedrijven-2015/21/overzicht

ENISA (2015), “ENISA Threat Landscape 2015”, [online], https://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-threat-landscape/etl2015

Raugas, M. et al. (2013), "Cyber V@R - A Cyber Security Model for Value at Risk", [online], https://www.cyberpointllc.com/docs/CyberVaR.pdf

Statistics Netherlands (2014), "Approaches of domestic product (GDP); National Accounts", [online], http://statline.cbs.nl/statweb/publication/?vw=t&dm=slen&pa=82262eng&d1=0-4,9-17,20-21,88,91,94,97,130-132,135-136,139,142&d2=(l-7)-l&hd=160105-1623&la=en&hdr=g1&stb=t

Statistics Netherlands et al. (2015), "ICT, Kennis en Economie 2015", [online], http://download.cbs.nl/pdf/ict-kennis-economie-2015-pub.pdf

Tetlock, P.E. and Gardner, D (2015), *Superforecasting: The Art and Science of Prediction*, The Crown Publishing Group

Van Wieren et al. (2016), *Understanding Bifurcation of Slow versus Fast Cyber Attackers*, submitted

Verizon (2015), "2015 Data Breach Investigations Report", [online], https://msisac.cisecurity.org/whitepaper/documents/1.pdf

World Economic Forum (2015), "Partnering for Cyber Resilience. Towards the Quantification of Cyber Threats", [online], http://www3.weforum.org/docs/WEFUSA_QuantificationofCyberThreats_Report2015.pdf

Reflexive Control in Cyber Space

Margarita Levin Jaitner[1] and Harry Kantola[2]
[1]Swedish Defence University, Stockholm, Sweden
[2]Finnish National Defence University, Helsinki, Finland
Margarita.jaitner@fhs.se
Harry.kantola@mil.fi

Abstract: Imagine it is possible to make an opponent make a decision advantageous to you by the simple use of information. According to Russian methodologies, such a capability exists and is known as the theory of Reflexive Control or RC. The theory itself has been researched in the Soviet Union and Russia since the 1960s and encompasses a methodology where specifically prepared information is conveyed to an opponent, which would lead him to make a decision desired by the initiator. The methodology is assumed to be applicable in a wide variety of situations, and is deeply rooted within the Russian Information Warfare concepts. Because theory envelops the Russian understanding of information as both technical data and cognitive content, "information resources" are understood as technological as well as human. Thus, RC can extend over the full range of machine, human and their interaction. The essential tasks of RC are for one the tracing of behavioural patterns within the system and for the other the identification "weak" information resource within the system. Arguably, a well-developed (global) cyberspace presents theorists and operators of RC and RC-methodology with a vast amount of possibilities to affect their opponent. This article explores ways in which RC can be exercised with the help of the cyberspace. In essence, the article shows that the cyberspace significantly enhances the potential of RC because it allows the initiator of RC to compromise both the system and its operator. It is then suggested, that RC is exercised to its full effect when the operator's cognitive assessment and the data residing in the system are compromised simultaneously.

Keywords: reflexive control, cyber space, information warfare, influence operations

1. Manipulating decision-making

> *"The quality of decision is like the well-timed swoop of a falcon which enables it to strike and destroy its victim."Sun Tzu*

In a struggle, one of the foremost tasks is to interfere with the adversary's decision-making process. One of the simplest taxonomies for decision-making processes is the subdivision into human-only, machine-only automated as well as human machine-assisted and collaborated decision-making systems.

In current military decision-making process, the human machine-assistant process is most prevalent. Machine-only automated decision-making systems are currently still frowned upon (Kott, 2015) and while machines may be taking over more and more steps of the process, it is not likely that humans will disappear as decision-makers any time soon. Therefore, the presented paper will focus primarily on collaborative and machine-assisted decision-making systems. Two distinctive potential attack points can be identified in an environment of a human machine-assisted decision-making process. For one, the adversary can try to influence the human and for the other the machine.

Modern decision-making processes emphasize the importance of recurrent gathering and evaluating of information, a comprehensive approach, in order to create courses of enemy as well as own action (COAs). In this way, COAs are for the most part based on intelligence and information provided by various situational awareness (SA) systems, weapons systems and the like. Thus, decision-making processes are heavily reliant on collection of data that is purposeful, correct as well as timely. Inaccurate and/or irrelevant information as well as delay in presentation can seriously cripple a decision-making process. In our context of a human machine-assisted decision-making this means that false, irrelevant or untimely information can be introduced to the human or the machine or both.

Arguably, decision-making processes follow certain patterns, or logic. Such patterns can reside on various levels and may be technological as well as human. Within technological systems, the range stretches from simple warning systems that are set off by a pre-programmed value, such as a conventional smoke detector, to complex systems that take a multitude of factors in account. In human decision-making process, the patterns are constructed through human behaviour on individual as well as group level. Such patterns are constructed through scores of factors ranging from cultural aspects, organizational structure to individual characteristics, such as propensity to take risks amongst the decision makers. Of course, in many cases, the complexity of such

patterns rises alongside with criticalness of the decision-making system. Thus, mapping of decision-making patterns may present an extremely challenging, but still achievable task. It is the knowledge of patterns within the decision-making process that allows an adversary to insert information into the process that would ultimately allow manipulating the decision.

The aim of this paper is to explore how the theory of Reflexive Control can assist gaining an advantage over an adversary's decision-making process. First, we will provide an overview over the theory of Reflexive Control and exemplify its general use based on events in the near past. Then we will describe three scenarios for manipulation of data. Each scenario will be covered from two perspectives – a cyber perspective as well as a cognitive-informational perspective. Following these scenarios, we will present a fictive case, where cyber as well as cognitive-informational manipulations are applied based on the theory of Reflexive Control. Finally, we will provide conclusions on the usability of RC in the context of Cyber and Information Security.

2. Theory of reflexive control

The theory of Reflexive Control (RC) originated in the Soviet Union in the 1960 and has been more-or-less continuously developed ever since, with some of the original researchers in the area still actively conducting research in the area. Amongst the scientific fathers of the theory are the V. A. Lefebvre, who now resides and works in the US, V. E. Lepsky, associated with the journal "Reflexive processes and control" as well as the resource reflexion.ru, but also M. D. Ionov and S. Leonenko. The approach takes background in ideas established in the eastern hemisphere such as the strategic thinking of Sun Tzu and particularly the Chinese use of stratagems. For example, Niu, Li and Xu's (2000) emphasize ten stratagems to be used in Information Operations.

In general, RC can be defined as "a means of conveying to a partner or an opponent specially prepared information to incline him to voluntarily make the predetermined decision desired by the initiator of the action"(Thomas, 2004).

The essence of a "reflexive game" according to Lefebvre is the mutual attempts of the adversaries to impose RC over one another (Kramer et. Al., 2003). This requires the adversaries to analyse their own as well as the adversary's ideas, and to model the adversary's behaviour in accordance (Kramer et. Al., 2003 and Thomas, 2004). Therein, reflexivity stands for the ability to create a correct model (Thomas, 2004). The more accurate the model, the more precise will be the prediction of the adversary's behaviour, the better will be the ability to introduce the desired "information package" to the adversary.

In a military context, M.D. Ionov (1995) identifies four distinctive ways to introduce such an "information package" to the adversary:

- Pressure by show of force. Such show of force can be exercised in various forms stretching across different aspects from diplomatic or economic pressure, such as threat of economic sanctions, to threats of military action, such as increasing combat readiness of armed forces or provocation, to declaration of war.
- Providing false information. This approach suggests the use of maskirovka – camouflage, denial and deception – on all levels in order to manipulate the adversary's perception of situation. This includes showing great force where there is indeed a weakness and vice versa, as well as the use of Trojan horse techniques.
- Affecting the adversary's decision-making process. Such approach includes systematic modelling of processes, publication of deliberately distorted doctrines, as well as presenting false information to the adversary's system and key figures.
- Affecting the timing of decision. Here, the element of surprise might be employed by sudden beginning of military operation or misleading the adversary to focus on another area of conflict to delay his reaction.

The term "information" should be understood in a broad fashion as it also includes emotional and a controlling elements. Show of (military) force, for example, may not so much aim at presenting the size and equipment of troops, but serve to intimidate, to provoke an emotional reaction. At the same time, the information can also be introduced at machine level. (Thomas, 2004)

Identifying the weak link within the adversary's processes, the point at which the "information package" can be introduced is central to RC. Likewise, it is necessary to know what type of information needs to be included into

the package. As Leonenko (1995) describes it, "RC consists of transmitting motives and grounds from the controlling entity to the controlled system that stimulate the desired decision. The goal of RC is to prompt the enemy to make a decision unfavourable to him. Naturally, one must have an idea about how he thinks."

3. Manipulation of data

Numerous ways to manipulate cognitive information and data exist. These range from withholding information or access to it, to providing the opponent with false information to create a deception and flooding the opponent with information of varying significance.

Withholding information from the opponent is often a desirable technique to conceal strength, weakness as well as immediate and even long-term plans. A general distinction can be made as to whether the opponent is aware of information being withheld. Particularly in those cases, where the opponent is unaware of an important fact missing, their situational awareness becomes incomplete. Own awareness of the opponent's incomplete assessment can then be exploited for own advantage.

In those cases where the opponent is aware of, or at least suspects the existence of the withheld information, the action of concealment may be straightforward. In a more intricate approach, the fact that the information is hidden would be emphasized. This would force the opponent to focus efforts on uncovering the missing bit of the puzzle, potentially sending him on a wild goose chase.

Leonenko describes the use of computers as a possible hinder, since computer power can be used to calculate, model and simulate situations and thus reveal the reflexive control attempt (Thomas, 2009). However, in some cases the opposite may as well be the truth. Withholding data from information systems used to assist decision-making will affect the results that these systems return. The aim can range from bluntly disabling the decision support system to covertly forcing the system to present the opponent with false values; first case, the decision-maker would have to take actions based on insufficient information, potentially making decisions that are favourable for the opponent. In the latter case, where the manipulation is desired to stay covert, sound knowledge of the information system is required in order to be able to carefully choose information to withhold. Prior information operations may be needed in order to make the value returned by the system believable to the decision maker. Even in those cases, where the value returned by the system appears incorrect to the human decision maker, the opponent may gain an advantage by simply sowing distrust and confusion, which in particular would hamper with decision-making in a fast-paced environment. (Kott, 2015)

In cyberspace, withholding information can for example through interfering with communication means (Chatham House). For instance, satellites are relatively vulnerable to external actions, since after their launch it is impossible to update the hardware while software updates are restrained by the overall architecture (Santamarta, 2014 and Hackett 2015). Activities may also include destroying information or parts of it in data centres, redirecting information searches to faulty sites, or overloading the system so that access can not be created, i.e. via different versions of (R)DDoS attacks, or disabling sensors. Of course, decision-making in military context are mostly done in closed circuit networks or in military specific environments, which are often though to get access to via cyber space. This does not mean that it is impossible, but it requires substantial amount of planning and preparations, in addition to actually gain physical access to the network (Zetter, 2015).

Information overload as a method is directly opposite to withholding information and "…occurs when information received becomes a hindrance rather than a help when the information is potentially useful" (Bawden et al., 1995). On the cognitive level such an approach amounts to presenting the opponent with masses of information. Knowledge of the opponent allows tailoring loads of information in a way that would catch his attention. This way, decision-making can be stalled by the time the opponent needs to process or dismiss the information. Thus, the information must be of some potential value, or it could simply be ignored. It must also be accessible, or the overload will remain potential, not actual. (Bawden and Robinson, 2009)

Altering information can appear to be a more complex technique than withholding information. The benefit of this approach, however, is that it is often hard to detect. Gradually changing information, it will eventually alter the outcome of the analysis and thus direct the opponent actions towards the desired end state. (Kantola and Hämäläinen, 2013) This applies at cognitive level, such as information directed at decision-makers through open source channels, as well as at machine level.

Altering information also has a higher reliability and predictability of how the opponent will act. The challenge lies in the need to have a high-level insight into the system in question, aside from having access to it. Sufficient insight would include a technical understanding of how data is being processed within the system. Further, a cultural and organizational understanding of how the human decision-makers analyse and interpret the values returned by the system is required. As in previously presented cases, prior operations might be necessary to "prepare" the decision-maker to accept the values that the affected system would return. Whereas a sudden change in results presented by the system, or when the system would present values that are inconsistent with the decision-maker's perception of the situation, the manipulation is likely to be quickly discovered. Similarly to previous examples, distrust and confusion is likely to ensue. The level of distrust and confusion is highly dependent on the characteristics of the human decision-makers – culture, overall trust in technology, existence of contingency plans and other factors. Studies have shown that inaccurate information may present a greater problem than information denial, because inaccurate information affects the preconception of the situational awareness. (Bryant and Smith, 2013)

Increased use of Blue-force Tracking-type[1] functionalities provides opportunities for efficient utilization of information altering technique. (Bryant and Smith, 2013) Force tracking systems usually present both blue (friendly) and red (hostile) force information retrieved by manual entry as well as updates by a network of different sensors. This provides the opponent with a multitude of attack vectors to utilize reflexively. (Thurston et al. 2013) Blue-force-tracking systems often uses a standardized protocol for exchanging information, which makes it easier for an attacker to understand how to alter the information. Furthermore, such systems utilize an increasing number of various [weapon] sensors, (Thurston et al, 2013), which multiplies ways to alter information provided to the decision-maker.

False or "properly modified" information, including contradicting information, can be provided to the opponent via the cyber space directly for the decision-maker's use, if a channel is established to their situational awareness system or placed "available" in suitable forums, databases or information sources for pick-up into their own network.[2] As previously, the values returned to the decision-maker should make sense to him, thus it may be required to simultaneously provide supporting information through other channels. Similarly, compromising only one sensor system would in most cases remain insufficient.

4. Case

In the previous chapter, we have introduced a number of techniques that can be used for manipulation of information on cognitive and machine level following the principles of RC theory. We will proceed by demonstrating how RC can be applied in a fictional operational setting. In our scenario, neighbouring countries A and B are entangled in a dispute. Country B possesses significant cyber capabilities and has a tradition of applying RC in politics as well as on the battlefield. Both countries are aware of three strategically important areas on the territory of Country A – X, Y and Z, which Country B assesses to be desired points of attack. (See Fig 1.) Country B initiates an operation in 5 phases during which it will attempt to exercise reflexive control over Country A's military and political decision-makers.

Phase 1. As the tensions between the two countries rise, Country B begins to prepare a military resolution of the conflict. B begins to mobilize troops and conducts exercises that can be perceived as targeting Area X. Country A's intelligence intercepts a number of leaked documents that point to imminent offensive with Area X as a target. Country A's CERT and Cyber Defence Units also identify a significant rise in relatively small-scale attacks, which they can with relative probability link to hacktivist and cybercriminals with ties to B. Country B's media reports on mobilization, which Country A's reporters quickly pick up, which probably results in increasingly hostile chatter between the countries in social media.

Phase 2. The next phase of B's operation begins as documents are planted for A's intelligence service to find within the service's regular activity. The documents suggest that Areas X and Y are less suitable for an initial offensive from B's perspective. The discovery of these documents coincides with what A perceives as B

[1] Blue-force Tracking is a United States military term and stands for GPS-enabled automated tracking of friendly forces.

[2] Placing information in a honeypot or honey net can be used for this purpose. This is a modified technique of responsive cyber defence derivated to suit delivery of information. This technique requires first active hacking to establish the channel, after which techniques described by Maybaum & Stinissen can be used. Technological techniques are described in Maybaum. M and Stinissen J. (2013) Responsive Cyber Defense; Technical and Legal Analysis. NATO CCD COE, Tallinn

deescalating at the highest political level. While B's troops remain mobilized A's political level attempts to engage in political resolution.

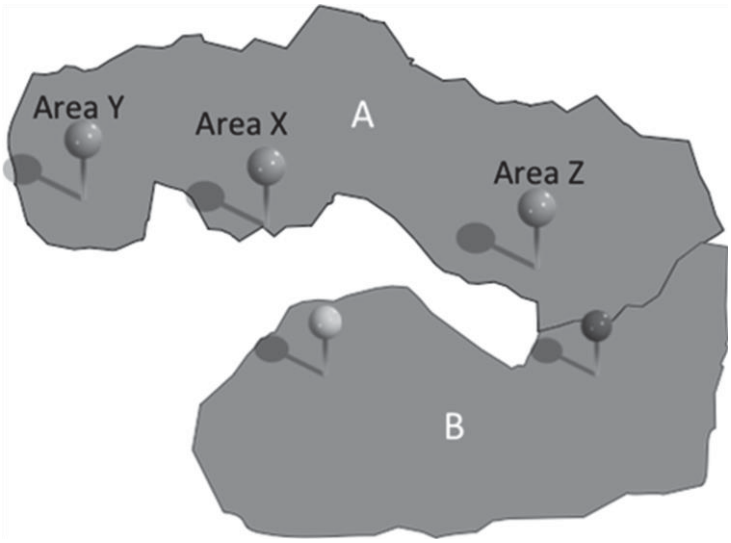


Figure 1: Generic scenario of countries A and B with strategically significant areas Y, X, Z identified on A's territory

Phase 3. The third phase is initiated by B's sudden exercise that implies an imminent attack on Area Y. The exercise is widely communicated to media and B actively encourages debate in social media regarding the details using "trolls"/opinion agents. At the same time B abruptly terminates diplomatic de-escalation. A's CERT and Cyber Defence Units identify increased small-scale cyber activity targeting the country's governmental and privately owned systems. A extends the high-alert level for troops despite the earlier indication that Y (as well as X) are less advantageous for B's offensive.

Phase 4. A brief period of de-escalation follows, during which B makes sure that the previously leaked documents regarding Area X and Y's unsuitability for B's offensive become widely known to A's civil society. For this purpose B may want to activate analysts, that have a history of supporting and propagating B's views in A's media. At this point, B utilizes any resource, that may encourage A's society to terminate the state of high-alert.

Phase 5. In the final phase of the operation, B attempts to give a credible impression of imminent attack against area Z. Country B choses to simultaneously spoof A's military and civilian air surveillance systems, creating an impression of planes flying into certain bases adjacent to Area Z. Simultaneously B lets information surface regarding ground transports in the same direction. The aim is to create a perception of massive troop movement into that area – which also corresponds with the plans initially leaked during the first de-escalation in phase 2. Ideally, this operation does not remain virtual, but is supported by factual movement of transportation trucks and trains towards the specified area in order to avoid discovery of the deception. In case A's cyber intelligence is known to be able to fetch information, fake plans and information could be planted for the intelligence to be gathered or, if possible, even planted straight into their systems.

Cyber activities have to be conducted both towards selected sensors, e.g. the air surveillance system, as well as supporting information sources, e.g. databases. Coordination with media reporting and other (planted) evidence is highly desirable, if not necessary. Social media offers a range of possibilities – troops available along the presumed transportation route can be encouraged to post pictures, geo-tag activities and the like. Bots then can multiply this human activity in social media.

At this point, the previously provided false information of Area Z being the most favourable point of attack and the situational picture provided by sensors and systems for decision-making assistance converge. Public appearance, such as in media and social media, confirms the picture to A's decision makers. At this point the A's conviction of an imminent attack in area Z should be of such confidence, that it triggers a defensive operation towards the area. At this point, the defensive operation at Area Z is likely to be supported within the society.

The phase culminates as B launches a factual offensive against area Y, surprising A. Preparations for this activity would have happened during the previous phases on a low scale – from initial mobilization of troops in the initial phase, to preparing the actual offensive during phase 3.

Activities in the different phases can be summarized as follows:

Table 1: Informational activities during the 5 Phases of the generic reflexive control operation

Phase	Actor B			Actor A
		Cognitive-Informational activities	Cyber activities	
0 Initial stage	Actors A and B are in discord, Actor B decides to resolve discord militarily, 3 possible attack areas identified within Actor A's territory – X, Y Z			
1 Pressure and mobilization	Mobilization of troops	Pressure against point X is promoted in media as well as through leaked documents	Low key supporting activity conducted by "hacktivists" and cybercriminals	Threat perception at X, high alert with focus on X
2 Planned planting and leaking of information	De-escalation	Leaking of documents suggesting X and Y less suitable for attack, Z preferred	Planting of documents suggesting X and Y less suitable for attack, Z preferred	Attempts to engage in diplomatic resolution, acute threat perception eases
3 Change of focus area	Engagement in diplomatic resolution halts abruptly, Maneuvers and simulations of attack against area Y	Pressure against point Y is promoted in media as well as in social media	Low key supporting activity for the information campaign, employment of "hacktivists" and cybercriminals	Tension rises, close observation, possibly inconclusive intelligence analysis
4 Creating confusion	Troops remain covertly mobilized towards Y, otherwise: de-escalation	Full scale promotion of area Z's higher suitability for attack than X, Y	Low key supporting activity for the information campaign	Inconclusive intelligence analysis
5 a. Major deception	Hide further mobilization towards B in the "noise"	Clear and massive indication of troop movement towards Z in media & social media	Attack sensors supporting perception of troop movement towards Z	Sensor results support intelligence analysis. Troop concentration towards Z.
5 b. Attack	Attack at Area Y			

According to the scenario, Country B applies the following methods in order to deliver the required information to Country A's decision-makers:

- False information is been pushed into the opponent's systems and/or provided to be fetched by A's intelligence gathering routine – Phases 1, 2, 3, 5
- Information being been altered in third-party/open systems – Phases 5
- Manipulated information is being provided to sensors – Phase 5
- Accurate information is concealed – Phases 1 – 5
- Mass-dissemination of information – Phase 1 – 5

An important element within the operation is the control over the consistency of overall information available to A's decision-makers and society. This allows Country B to create confusion, when the purposefully leaked false assessments do not correspond with actual mobilizations, in particularly in Phase 3. This also allows Country B to evoke Country A's confidence in its situational assessment in the final phase, Phase 5. A confidence then leads to action based on the holistically manipulated situational picture. Thus, during the course of operation,

Country B establishes relative information superiority, which allows it, to a certain degree, to steer its opponent's reactions reflexively.

As previously noted, the scenario is generic and aims to demonstrate as many of the manipulation techniques as possible. The variation of techniques that can be applied in an actual scenario may differ significantly as it depends upon numerous factors. Technological abilities, cyber readiness and level of skill of cyber operation planners, but also knowledge of the opponent are amongst these factors. Ultimately, each real situation will be unique and will require an individual solution.

5. Conclusions

In the presented paper we have described how manipulation of cognitive information and cyber operations can be manipulated with the aim of gaining an advantage over an opponent. We have also demonstrated how it is possible to use both, (cognitive) information warfare and cyber warfare methods for operations that follow the principles of RC. In particular, our focus lied on the advantage of combining information warfare – as warfare on the cognitive level – with cyber means, thus manipulating the situational awareness of a modern commander who seeks support in technical solutions for decision-making.

The paper shows that it is possible to use both information warfare and cyber warfare methods to operate according to the principles of reflexive control. More so, it appears to be of advantage to combine information and/or psychological operations in a combined manner when attempting to create a deception regardless application of RC or any other theory or technique.

We have also demonstrated that operations based on RC theory are designed as longer-term operations. They require an intimate understanding of the opponent, his foundations and reactions to various information and stimuli. Likewise, such operations draw advantage of a longer engagement with the opponent aiming at step-by-step preparing him to take a decision predetermined by the initiator of RC. An aggravating factor within is that no exact operation can be used twice, as it would open up for the opponent's own learning, making the initiator himself vulnerable to RC.

Nevertheless, we can argue that such operations can create decisive outcomes and thus be worth the struggle. The study also shows that there is an advantage of using cognitive (InfoOps/PsyOps) operations together with Cyber operations. Actions taken in the spheres of cognitive information and cyber have to be well planned, prepared and coordinated. Trustworthiness of information is key to their success.

Because this paper carries an exploratory character, we strongly suggest further academic research of the individual elements of cooperation in the borderland between cyber and cognitive information.

References

Bawden D, Holtham C, Courtney N. (1995) Perspectives on information overload. Aslib Proc New Inf Perspect 1999;51(8):249–55.

Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. Journal Of Information Science, 35(2), 180-191. doi:10.1177/0165551508095781

Bryant, D. J., & Smith, D. G. (2013). Impact of Blue Force Tracking on Combat Identification Judgments. Human Factors, 55 (1), 75-89. doi:10.1177/0018720812451595

Chatham House, Cyber and Space, online http://www.unidir.ch/files/conferences/pdfs/a-review-of-the-chatham-house-space-and-cyber-linkages-project-en-1-983.pdf (last accessed 2016-02-07)

Hackett, R., (2015) Here's the scary new target hackers are going after. Fortune online, http://fortune.com/2015/08/04/globalstar-gps-satellite-network-hackers/ (last accessed 2016-02-02)

Ionov M.D. (1995) On Reflexive control of the Enemy in Combat // Military thought (English edition) No.1 (January 1995), pp. 46,47.

Kantola, H., & Hämäläinen, J. (2013). Modelling Cyber Warfare as a Hierarchic Error Effect of Information. Proceedings of the 12th European Conference on Information Warfare and Security: ECIW 2013 (p. 322). Academic Conferences Limited.

Kott, A., (2015) War of 2050: a Battle for Information, Communications, and Computer Security, US Army Research Laboratory

Kramer, X.H., Kaiser, T.B., Schmidt, S. E., Davidson, J. E., Lefebvre V. A. (2003) From Prediction to Reflexive Control. – Reflexive processes and Control. – Nr. 2. – Vol 3. p. 35.

Leonenko. (1995) Refleksivnoe upravlenie protivnikom (Reflexive Control of the Enemy), Armeyskiy Sbornik, No 8. 1995 p. 28.

Lepsky, V.A., Stepanov A.M. (2003). Peculiarities of Reflexive Processes in Religious Worship Organizations. Reflexive Processes and Control. No. 1. Vol. 2.

Maybaum. M and Stinissen J. (2013) Responsive Cyber Defense; Technical and Legal Analysis. NATO CCD COE, Tallinn

Niu L., Li J., and Xu D., (2000) Planning and Application of Strategies of Information Operations in High-Tech Local War, Zhongguo Junshi Kexue (China Military Science), no.4, pp.115–122.

Santamarta, R., (2014), A Wake-up Call for SATCOM Security, Technical White Paper, IOActive, http://www.ioactive.com/pdfs/IOActive_SATCOM_Security_WhitePaper.pdf (last accessed 2016-02-07)

Thomas, T.,(2004), Russian Reflexive Control, Theory and the Military, Journal of Slavic Military Studies 17: 237–256, ISSN:1351-8046.

Thomas, T. (2009). Nation-state Cyber Strategies: Examples from China and Russia. Cyberpower and National Security, 475-76.

Thurston, M., Stephens, B., Daniels, M. R., & Steinberger, J. (2013). Building a Culture of Efficiency in Blue Force Tracking Technology. Defense AT&L, 42(5), 12-16.

Zetter, K., (2015), Researchers hack air-gapped computer with simple cell phone, www.Wired.com, http://www.wired.com/2015/07/researchers-hack-air-gapped-computer-simple-cell-phone/ (last accessed 2016-02-02)

Automating Cyber Defence Responses Using Attack-Defence Trees and Game Theory

Ravi Jhawar[1], Sjouke Mauw[1] and Irfan Zakiuddin[2]
[1]Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg
[2]Noumena Research Ventures Ltd., UK
ravi.jhawar@uni.lu
sjouke.mauw@uni.lu
noumena.research.ventures@gmail.com

Abstract: Cyber systems that serve government and military organizations must cope with unique threats and powerful adversaries. In this context, one must assume that attackers are continuously engaged in offence and an attack can potentially escalate in a compromised system. This paper proposes an approach to generate defensive responses against on-going attacks. We use Attack-Defence Trees (ADTrees) to represent situational information including the state of the system, potential attacks and defences, and the interdependencies between them. Currently, ADTrees do not support automated response generation. To this end, we develop a game-theoretic approach to calculate defensive responses and implement our approach using the Game Theory Explorer (GTE). In our games, Attackers and Defenders are the players, the pay-offs model the benefit to each player for a given course of action, and the game's equilibria is the optimal course of action for each player. Finally, given the dynamic nature of cyber systems, we keep our ADTrees and the corresponding game trees up-to-date following the well-known OODA (observe, orient, decide, act) loop methodology.

Keywords: cyber defences, attack modelling, game theory, security, incident response

1. Introduction

Cyber systems are becoming highly complex with ever-increasing dependencies both internally as well as with strategic partners and commercial service providers. Military organizations and critical businesses are also relying heavily on such cyber systems to meet their operational demands and to support mission execution. At the same time, cyber attacks are becoming stealthy and sophisticated, posing potentially very high damaging impact. In this context, a holistic framework for responding to cyber attacks becomes essential and it must encompass several functions including:

- efficient collection of cyber situational information,
- analysis of possible attacks,
- determining the courses of actions in response, and
- taking the appropriate actions.

This paper focuses mainly on the 'determining the courses of actions in response' component of a deployed cyber system. We assume that situational information including the system state and parameters, and attack and defence related information is available. In this work, we systematically represent the situational information using Attack-Defence Trees (ADTrees) (Kordy, Mauw, & Radomirovic, Attack-defence trees, 2014). ADTrees improve the widely used attack trees formalism, by including not only the actions of an attacker, but also possible counteractions of a defender. The root node in an ADTree represents the attacker's (or defender's) goal and the children of a given node represents its refinement into sub-goals. Each node can have one child of the opposite type, representing the node's counteraction, which can be refined and countered again. The leaves of an ADTree represent the basic actions of an agent, which need not be refined any further.

Formally, ADTrees extend the formalism of defence trees (Bistarelli, Fioravanti, & Peretti, 2006), where defensive measures are not refined and can only be attached to leaf nodes. ADTrees can also be seen as merging attack trees and protection trees (Edge, Dalton, Raines, & Mills, 2006) into one formalism. Protection trees are AND-OR trees depicting how defensive measures can be refined into simple actions. Given the high expressivity and intuitiveness of ADTrees, complemented with strong mathematical foundations, they seem as an appropriate choice to describe and analyze cyber situational information.

Currently, ADTrees are used to analyse and quantitatively assess security scenarios. They do not compute the course of actions as responses against on-going attacks. We propose to address this limitation by applying game theory to ADTrees. Game theory provides a rich resource of mathematical and algorithmic tools to study the problems of competition or conflict. We view a *cyber response problem* as a game between an attacker who is competing to inflict some form of attack and a defender who is attempting to prevent the attack. A game-solver then computes the best responses to defend the cyber system from various attacks launched by an attacker.

Kordy et al. in (Kordy, Mauw, Melissen, & Schweitzer, 2010) have already established a two-way mapping and equivalence between games and ADTrees. However, they consider games of a highly restricted form that are not suitable in our context because of the following limitations:

- They use only binary pay-offs; this implies that there are only two possible outcomes: the attacker wins and the defender loses, and vice versa.
- They assume existence of perfect information implying that both players have full knowledge of all opponent's actions.
- Strict alternation of player's moves is required. Assuming an alteration between the attacker's and the defender's moves may be unrealistic in our case.
- Finally, the mappings in (Kordy, Mauw, Melissen, & Schweitzer, 2010) result in an increased abstraction from reality. Each mapping consists in generating a suitable syntactic object called an ADTerm that maps the binary pay-offs of the game tree. The way in which such syntactic objects represent the real world is unclear.

In this paper, we address the limitations in (Kordy, Mauw, Melissen, & Schweitzer, 2010) and define a game model that has the capability to represent the cyber response problem. Section 2 presents a motivating scenario that places our work in context. Section 3 provides the fundamental definitions of our game model and Section 4 defines the mapping between ADTrees and the basic form of our game model. Section 5 then extends our basic game to allow modelling of complex cyber response problems. Section 6 defines our approach for updating game trees following the OODA loop and Section 7 outlines our conclusions.

2. Motivating scenario

Consider a military organization that has deployed a small, dedicated cyber system to support one of its missions. The mission might be for a Remotely Piloted Aircraft System (RPAS) to track an object in a geographical region. The cyber system performs functions like storing and processing the image and location data sent by the RPAS in order to generate the navigation plans (NP). This cyber system can also be a subnet separated by a firewall within a large distributed network operated by the military organization.

Assume that the dedicated cyber system consists of a file server (FS) which stores the image and location data, a navigation plan generator (NPG) that computes the future navigation routes for the RPAS, and three client workstations (WS) that control the RPAS. FS offers file transfer (ftp), remote shell (rsh) and secure shell (ssh) services to WS so that they can access the image and location data. NPG on the other hand allows WS and FS to execute commands on it using the ssh service. A firewall, which is intended to protect FS and NPG, only allows ftp, rsh and ssh traffic from WS to FS and NPG and blocks all other traffic. Let us further assume that there are vulnerabilities in ftp and ssh daemons, in the task scheduler of NPG, and in the address space resolution of FS's operating system. The access control list defines that a user has read and execute privileges while a root can read, write and execute. Finally, the goal of the attacker is to breach the integrity of the system so that the mission fails.

In ADTrees, attacks are represented as circles and defences as rectangles. Refinements are indicated by solid edges between nodes and counteractions are indicated by dotted edges. Attacks and defences can be refined conjunctively and disjunctively. A conjunctive refinement of a node has an arc connecting the edges going from this node to its children. A disjunctive refinement has simple edges.

The ADTree in Figure 1 shows how an attacker can modify critical mission data in two different ways and provides possible defence choices. In the first attack (see node "NPG root"), the attacker can obtain root privileges on NPG: first by gaining root privileges on WS via a key logging technique or by performing a buffer overflow attack on the ssh daemon. Then the attacker can use the interactive "cmd.exe" command. The defender can disable

the task scheduler to prevent the execution of the cmd.exe command; use a two-factor authentication scheme against the key logging attack; and stop the ssh service to prevent buffer overflow.

Figure 1: Example ADTree representing the attack-defence scenario for a military organization

In the second case, the attacker must obtain root privileges on FS (see node "FS root"). To achieve this, she must first gain user privileges on FS and then perform a local buffer overflow attack. The defender can prevent the latter attack by using adaptive memory management techniques. To obtain user privileges on FS, the attacker can, starting as a user on WS:

- exploit the ftp vulnerability and use the rsh service to establish trust between WS and FS, or
- perform a buffer overflow using the vulnerability in the ssh daemon.

The defender can prevent this attack by:

- modifying the access control list, or
- configuring the firewall to drop ftp packets from WS and blocking the rsh service, and
- stopping the ssh service.

3. Game model

We use ADTrees to express situational information and to analyze attacks in the system. However, we note that ADTrees are only a language to describe and formalize attacks and defences; they do not compute responses against attacks by themselves. We propose to solve the cyber response problem by applying game theory on ADTrees. We define a game between an Attacker who is competing to inflict some form of attack and a Defender who is attempting to prevent the attack. A game-solver then provides the cyber responses to defend the system from the attacker.

In this section, we define the following basic components of our games: the game's players, knowledge states of the players, game's moves, and the pay-off function. The next section discusses an approach to generate our games from ADTrees.

3.1 Game's players

Our model considers interaction between two players: a Defender and an Attacker. In general, considering a single Defender-player implicitly assumes that the Defender has nearly full knowledge of the state of the system and that she can implement any determined course of actions effectively. However, in comparison to centralized

control, a model with localized decision-making seems sensible. To this end, one approach consists in defining multi-player games with a number of Defender-players. Another approach consists in coordinating several two-player games of a Defender against the Attacker, where each game makes a localized decision, and the overall solution is the composition of local results. We adopt the latter approach by defining two-player local games focusing on specific locations or critical resources in the system (e.g., a local game where the Attacker attempts to gain root privileges on FS and the Defender aims to protect FS). We choose the latter approach because it allows us to define game models for each resource in the system and to take into account the distinct trust levels associated with each resource. For example, a cross-boundary located server has lower trust than an on-site server. Our game model addresses this aspect by adjusting the pay-off values based on the `risk appetite' parameter.

The Attacker attempts to breach the defences of the system in order to disrupt missions. In our example, the goal of the Attacker is to breach the integrity of a mission either by compromising FS or NPG. Assuming a single attacker implies that she has full knowledge of possible attack strategies and has centralized control for inflicting her actions on the system. Therefore, having a single Attacker-player provides a model with a very strong attacker and it may be desirable to retain this model, irrespective of the aforementioned models for the defender.

3.2 Game's moves and knowledge states of the players

When considering the game moves, we highlight a conceptual distinction between games and ADTrees, specifically regarding the notion of *strategy*. In a game, a strategy is a complete algorithm that tells a player what to do for every possible situation throughout the game. In ADTrees, concrete actions are only at the leaves and all other nodes define a refinement relationship using conjunction and disjunctive operators. Therefore, in an ADTree, the strategy is the connection between a concrete action and the goal that drives it.

For example, for the ADTree in Figure 1, an attacker can reach her goal following four *attack strategies*:

- a_1={WS user, ftp-rhosts & rsh, local-bof FS}
- a_2={WS user, sshd-bof, local-bof FS}
- a_3={WS key logger, interactive cmd.exe}
- a_4={ sshd-bof, interactive cmd.exe}

All actions within an attack strategy must be implemented to breach integrity, but implementing one of the four attack strategies is sufficient. While a_1 and a_2 allows the attacker to gain root privileges at FS, a_3 and a_4 compromises NPG. There are three *defence strategies* to protect FS:

- d_1={drop ftp or stop rsh, stop ssh}
- d_2={memory management}
- d_3={modify ACL}.

and two defence strategies to protect NPG:

- d_4={task scheduler} and
- d_5={2nd auth factor, stop ssh}.

Our goal here is to have the game's moves model the attack-defence strategies of each player.

The knowledge state of a player defines what the player knows of its moves and the moves of other players. The simplest case considers *perfect Information* where both players have full knowledge of all the moves of the game. In this work, we consider more complex models for knowledge states where the moves and pay-offs of both players are not fully known to either the Attacker or Defender. In particular, our game model allows moves where:

- The Defender cannot distinguish the choices made by an Attacker and vice versa.
- Game's moves are committed temporally independently and/or simultaneously.
- A player may choose not to take any action that changes the state of the system (e.g., a defender can simply monitor the network for possible intrusions).

- There is uncertainty in observations and expected pay-offs.

We believe that such complex game models can sufficiently represent the cyber security scenarios like the one described in Section 2.

3.3 Pay-off function

Pay-offs model the benefit to each player for a given course of actions/game's moves. In zero-sum games, the gains of one player are equivalent to the losses of the other. We note that pay-offs are also critical in capturing essential notions like the `risk appetite'. Assigning realistic pay-offs is a hard problem and is out of the scope of the work presented here. Instead, we start from a basic game model with arbitrary pay-offs where the pay-off to the Attacker is simply a measure of the amount of work the Defender has to do. Therefore, when the Defender has to take one action the Attacker receives a pay-off of 1. Intuitively, the goal of each player is to commit their game moves such that they maximize their own pay-offs.

4. Mapping attack-defence trees and games

In this section, we define a basic game and provide its solution. We then describe the notion of equilibrium and its relationship with the cyber response problem.

4.1 The basic game model and the game trees

To generate our basic game, we start from the ADTree that models the overall security of the system and compute all attack and defence strategies, as described in Section 3.2. In this section, we aim to model a two-player local game, focusing on a specific resource in the system (FS), as discussed in Section 3.1.

The players are clearly the Attacker and Defender; the game's moves that model the choices of each player are denoted as labels on the edges and take the form:

Player.InformationState.ActionType.

Player can take one of the two values *A* or *D* representing the Attacker and the Defender, respectively. *InformationState* is a number associated with each action representing the depth of the knowledge state of the player while making the game's move (see Section 3.2). In the simple example below, the decision for both players comes from their first information state. Finally, the *ActionType* refers to the concrete steps taken by each player – in our game, action types correspond to attack and defence strategies.

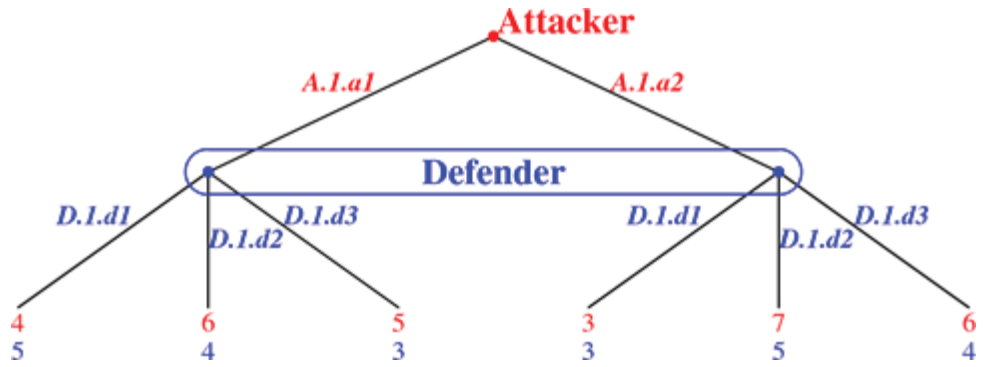


Figure 2: Basic game tree focussing on FS

For the local game focusing on FS, two attack strategies allow the Attacker to gain root privileges on FS:

- a_1={WS user, ftp-rhosts & rsh, local-bof FS}
- a_2={WS user, sshd-bof, local-bof FS}.

On the other hand, the Defender can protect FS – or respond to Attacker's moves – using one of the three defence strategies:

- d_1={drop ftp or stop rsh, stop ssh},
- d_2={memory management}, and

- d_3={modify ACL}.

For the sake of readability, action types in Figure 2 are denoted using the above attack and defence strategy identifiers. The Attacker must choose from a_1 and a_2 in order to implement its attack and, in response, the Defender must choose from d_1, d_2 or d_3.

A player's game strategy can be extracted by following a path from the apex to a leaf. The path presents the moves of both players. In Figure 2, the Defender circumscribes the two nodes that model the result of the Attacker's moves. We use this notation, following (Egesdal, et al., 2015) (Savani & Stengel, 2014), to denote the fact that the Defender cannot distinguish these nodes because it does not know the choice made by the Attacker. In other words, nodes circumscribed using a rounded rectangle are treated as a single information state for the Defender (see Section 5.2). Although the moves of the Defender are structurally presented as following the Attacker's moves, semantically there is no temporal dependency. This means that in our game model the moves are not assumed to be committed in any particular order and can even occur simultaneously.

The leaves of the game tree are the pay-offs and they measure the benefits to each player for their course of actions. For example, if the Defender chooses to strengthen the network by using memory management (strategy d_2) as a defence, although she can prevent the Attacker from gaining root privileges at FS, the number of tasks that the Defender performs is significant, thus the pay-off of 7 to the Attacker. We note that, although we have assigned the pay-offs here in a simple manner, we tried to take into account the risk appetite and the sensitivity of the impact of attack and defence choices.

4.2 Solving the games

We use a web-based game solver called the Game Theory Explorer (GTE) (Egesdal, et al., 2015) to obtain solutions for our games. The solution consists in computing the equilibria, which in our case describes the best game strategies for both players. Attacker and Defender are in equilibrium if the Attacker is choosing the best strategy she can, taking into account the Defender's strategy, while the Defender's decision remains unchanged. Similarly, the Defender is choosing the best strategy she can, taking into account the Attacker's decision, while the Attacker's decision remains unchanged.

Figure 3 illustrates a part of the solution of our basic game (Figure 2) as provided by the GTE. The 2x3 matrices characterize the pay-offs of both players (player 1 being the Attacker and player 2 the Defender). Each of the three rows EE1, EE2 and EE3 denote an equilibrium, with corresponding expected pay-offs.

```
Strategic form:

2 x 3 Payoff player 1                          2 x 3 Payoff player 2

        D.1.d1  D.1.d2  D.1.d3                          D.1.d1  D.1.d2  D.1.d3
A.1.a1       4       6       5              A.1.a1       5       4       3
A.1.a2       3       7       6              A.1.a2       3       5       4

EE = Extreme Equilibrium, EP = Expected Payoffs

Rational:
EE 1   P1: (1) 2/3 1/3 EP= 5    P2: (1) 1/2 1/2 0 EP= 13/3
EE 2   P1: (2)   1   0 EP= 4    P2: (2)   1   0 0 EP=    5
EE 3   P1: (3)   0   1 EP= 7    P2: (3)   0   1 0 EP=    5
```

Figure 3: Solution of the basic game

Let us look at the equilibria in detail.

- (Row 1) EE1 consists of player 1 (the Attacker *A*) playing a game strategy labelled (1) and this strategy is for her to make the first game move A.1.a_1 with probability 2/3 and the second game move A.1.a_2 with probability 1/3. As a response, player 2 (Defender *D*) can make game moves D.1.d_1 and D.1.d_2 with equal probabilities of 1/2 each, but does not play D.1.d_3. By following this game strategy, the Attacker can expect a pay-off of 5 and the Defender's expected gains are 13/3.

- (Row 2) The equilibrium EE2 consists of a game strategy labelled (2) where the Attacker only makes its first game move A.1.a_1 with probability 1. As a response, the Defender also makes only the game move D.1.d_1 with probability 1. The expected pay-offs through this game strategy is 4 for the Attacker and 5 for the Defender.
- (Row 3) The final equilibrium consists of game strategy (3) where the Attacker only makes its second game move A.1.a_2 with probability 1 and the Defender responds to it through the game move D.1.d_2 with probability 1. The expected pay-offs for the Attacker is 7 and for the Defender is 5.

We can use *the equilibria to identify the best defence responses against on-going attacks*. For example, in game strategy (2), when the Defender identifies that there is an *on-going attack* on FS following attack strategy a_1, the *best response* for the Defender is to drop ftp packets between WS and FS and to stop ssh service on FS (i.e., apply d_1). This allows the Defender not only to stop the on-going attack but also to strengthen her system without paying heavily in terms of the amount of Defender tasks required.

Our basic game model satisfies the 2nd and partially the 1st requirement listed in Section 3.2. However, to accommodate complex models – satisfying all four requirements – we need to extend our basic game so that our cyber response problem can be modelled holistically.

5. Extended game models

Our basic game model assumes a static scenario where the players consider all options upfront and make a strategy choice, which fixes a definite course of action for each player. The limitations of the basic model are:

- A player may choose not to take an action that changes the state of the system. For example, a Defender may only monitor the network to observe the situation and an Attacker may perform reconnaissance. Such *wait* conditions may be necessary, but were not included in the basic game model. To address this, we add a *wait* game move to the strategy sets of the players (see Figure 6).
- In networked systems, several unexpected system events and on-going attacks may go unnoticed. Since such observations are critical for successful execution of missions, we need to enhance our game trees and introduce probabilistic branching that takes into account the uncertainties about what has happened (see Section 5.1).
- In our basic game both players know the pay-offs to each other. This is an unrealistic assumption since it implies that each player knows the impact of a course of action on both itself and, critically, on its opponent. To address this limitation, we introduce randomization to capture variable pay-offs (see Section 5.2).

5.1 Randomization to capture uncertain observation

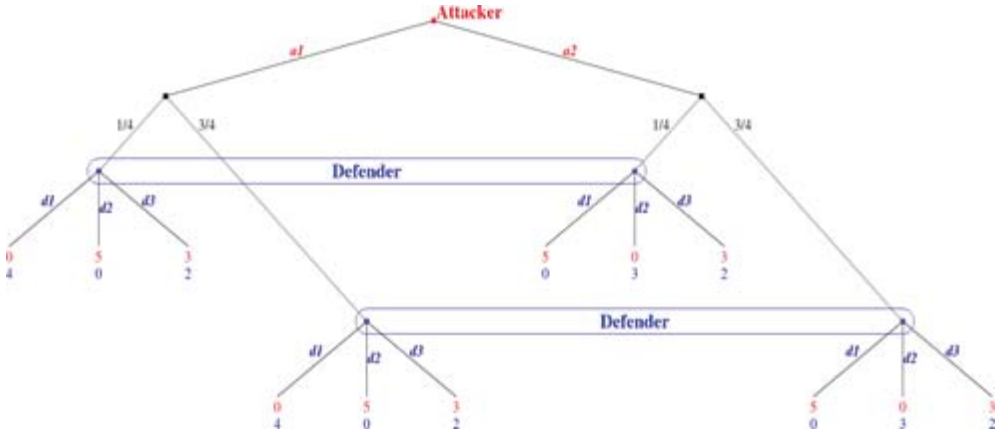


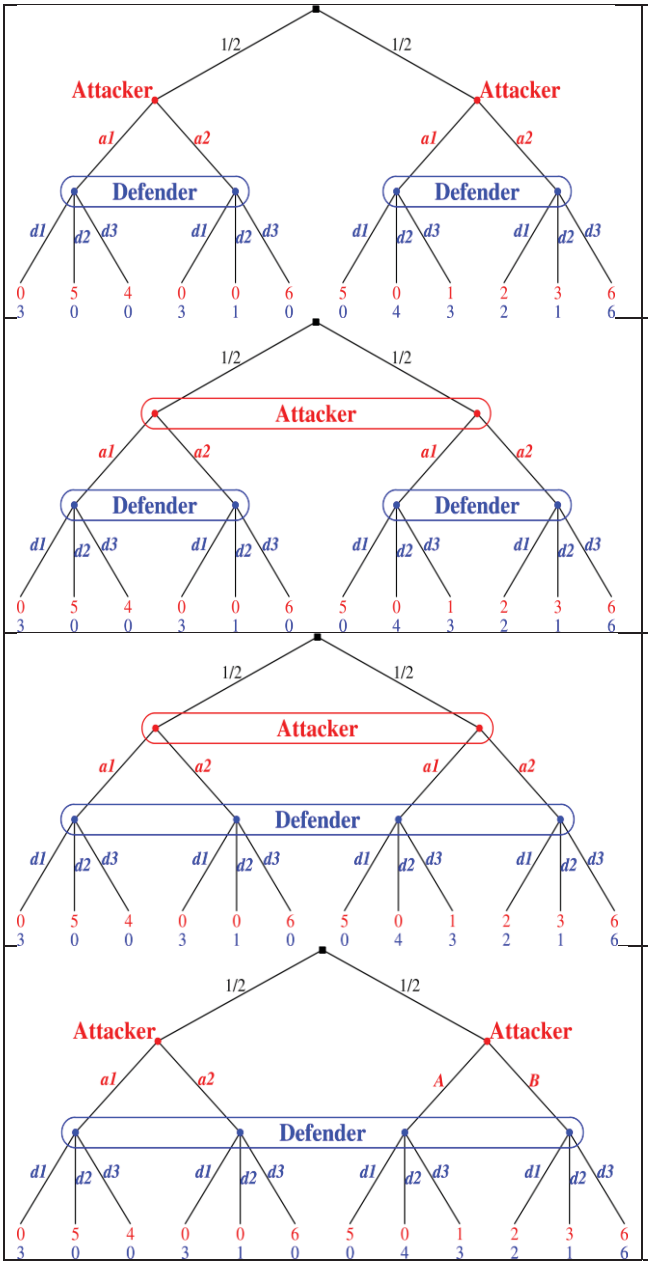
Figure 4: Game tree with uncertain observations

Figure 4 illustrates an example of how randomization in the game tree can model uncertainty of observation. The Attacker can choose either attack a1 or a2 (notation has been simplified here for readability). However, there is a probabilistic branch after her action, which leads to the choices for the Defender. Therefore, the Defender knows that there are 3/4 chances that the Attacker has committed a_1 and 1/4 chances of a_2 being played. In contrast, the Defender in our basic game did not know if Attacker plays a1 or a2.

5.2 Randomization to capture variable pay-offs

We consider four cases to introduce randomization of pay-offs. In the first game (Figure 5, row 1), there is an initial randomized branch with two sub-trees and corresponding pay-offs at the leaves. There are two possibilities of pay-offs and these are known to both players. As discussed in Section 4.1, observe that there is no rounded rectangle for the Attacker and two separate ones for the Defender, one for each case. In the second game (Figure 5, row 2), the Attacker does not know the pay-off possibility, but the Defender does, since we have an Attacker who is unable to distinguish the initial probabilistic branch. In the third game (Figure 5, row 3), neither the Attacker nor the Defender knows which pay-offs will be the case and in the final game, the Attacker knows which pay-off possibility will be the case, but the Defender does not.

We note that, although the same initial pay-offs are assigned to all the game trees, the solutions are very different (see col. 2 in Figure 5). The number and the kind of game strategies in each equilibria and expected final pay-offs for both the players vary significantly since their knowledge states change drastically in each game. Following our extended game models, we can generate a rich set of game trees that precisely represent the complex requirements of our scenario.




```
EE 1
P1: (1) 0 1 0                    EP= 7/2
P2: (1) 0 0 1 0 0 0 0 0 0        EP=
15/2
```



```
EE 1
P1: (1) 0 1                      EP= 3
P2: (1) 0 0 1 0 0 0 0 0 0        EP= 9/2
```



```
EE 1
P1: (1) 4/5 1/5                  EP= 5/2
P2: (1) 0 7/9 2/9                EP= 9/5

EE 2
P1: (2)   0   1                  EP=   6
P2: (2) 0   0   1                EP=   3

EE 3
P1: (3)   1   0                  EP= 5/2
P2: (3) 0   1   0                EP=   2
```



```
EE 1
P1: (1) 0 0 0 1                  EP= 6
P2: (1) 0 0 1                    EP= 3
```

Figure 5: Game trees with randomization to capture variable pay-offs and corresponding game solution

6. Deploying game models with the OODA loop

ADTrees capture cyber situational information in a static manner and support analysis of risks off-line. Cyber systems on the other hand are dynamic, with many system changes (e.g., migration of virtual machines, failure of storage disks) over time. Game trees are also a static formulation of interacting choices – a single tree cannot express the evolution of state over time. To address this issue, we propose to update our ADTrees and game trees in the events of system changes. We adopt the OODA loop methodology (Hightower, n.d.) as follows:

- *Observe* – collate information about cyber incidents and system changes.
- *Orient* – arrange collated information on suitable ADTrees.
- *Decide* – formulate games in concurrence with the updated ADTrees and solve them.
- *Act* – raise alerts to the system administrator with possible cyber response solutions to implement an appropriate action.

The loop reverts to the *Observe* step after *Act* and continues similarly thereafter. As an example, consider that at time instance *t*, it is *observed* that the Defender patches the ssh daemon and the Attacker is scanning for a new set of IP addresses in the network (*wait* game move). During *orientation*, attack a_1 is disabled and the corresponding defence d_1 need not "stop the ssh service" anymore (let d_1'={drop ftp or stop rsh}). Therefore, we update the game tree in Figure 2 and obtain the following game tree, which is then used to *Decide* about cyber responses.

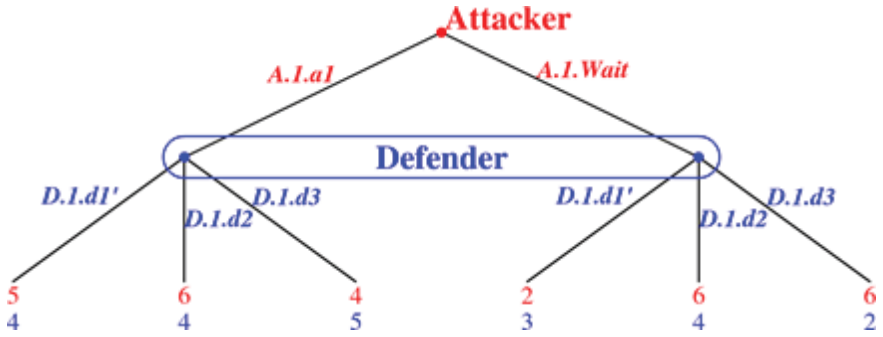


Figure 6: Updated game tree at time instance t

7. Conclusions

We proposed an approach to generate cyber defence responses by mapping situational information from ADTrees on to game trees. A variety of game models were demonstrated to support complex cyber response analysis, implemented by the GTE tool. Finally, we also account for dynamic system behavior by adapting our models following the OODA loop.

This work is an initial study that intends to understand the applicability of ADTrees and game theory in solving the problem of cyber responses generation. Our future work will first focus on defining a cost function that takes various functional and security parameters as input and provides the pay-off values as output. We will then focus on improving the scalability of our approach and perform experiments on realistic cyber testbeds.

Acknowledgements

The work of Sjouke Mauw and Ravi Jhawar was supported by the European Commission through the FP7 project TREsPASS (grant agreement n. 318003) and by the Fonds National de la Recherche Luxembourg through the ADT2P project (grant n. C13/IS/5809105).

Irfan Zakiuddin's contribution to this work was funded by the Defence Science and Technology Laboratory (DSTL), which is a part of the UK's Ministry of Defence. In DSTL, Irfan would like to thank Kevin Wise, the DSTL Technical Partner (TP), for another enjoyable project, much support and the usual string of great project meetings. Irfan would also like to thank Ed Moxon, the DSTL customer, for guidance and support. Irfan's work depended critically on huge amounts of free help from game theorists Rahul Savani and Theodore Turocy, both of whom repeatedly found time, amidst many more important commitments, to answer a barrage of questions.

References

Albanese, M., Jajodia, S., & Noel, S. (2012). Time-efficient and cost-effective network hardening using attack graphs. *DSN* (pp. 1--12). Boston, USA: IEEE.

Bistarelli, S., Fioravanti, F., & Peretti, P. (2006). Defence Trees for Economic Evaluation of Security Investments. *ARES* (pp. 416--423). Vienna, Austria: IEEE.

Edge, K., Dalton, G., Raines, R., & Mills, R. (2006). Using Attack and Protection Trees to Analyze Threats and Defences to Homeland Security. *MILCOM* (pp. 1--7). Washington, DC, USA: IEEE.

Egesdal, M., Gomez-Jordana, A., Pelissier, C., Prause, M., Savani, R., & Stengel, B. (2015). *Game Theory Explorer*. Retrieved from http://gte.csc.liv.ac.uk/gte/builder/

Hightower, T. (n.d.). *Boyd's OODA loop and how we use it*. Retrieved from Tactical Response: https://tacticalresponse.com/blogs/library/18649427-boyd-s-o-o-d-a-loop-and-how-we-use-it

Kordy, B., Mauw, S., & Radomirovic, S. (2014). Attack-defence trees. *Journal of Logic and Computation*, 24(1), pp. 55—87.

Kordy, B., Mauw, S., Melissen, M., & Schweitzer, P. (2010). Attack--Defence Trees and Two-Player Binary Zero-Sum Extensive Form Games Are Equivalent. *GameSec* (pp. 245--256). Springer.

Paul, S. (2014). Towards automating the construction & maintenance of attack trees: a feasibility study. *GraMSec* (pp. 31--46). Grenoble, France: EPTCS.

Savani, R., & Stengel, B. (2015). Game Theory Explorer: Software for the Applied Game Theorist. *Computational Management Science*, 12(1), pp.5—33.

Leadership for Cyber Security in Public-Private Relations

Tuija Kuusisto[1] and Rauno Kuusisto[2]
[1]Ministry of Finance & National Defence University, Helsinki, Finland
[2]The Finnish Defence Research Agency, Riihimäki, Finland & National Defence University, Helsinki, Finland & University of Jyväskylä, Jyväskylä, Finland
tuija.kuusisto@vm.fi
rauno.kuusisto@mil.fi

Abstract: Nation states published cyber security strategies few years ago. The reaching of the targets presented in the strategies often requires interaction and collaboration with the information and communication system and services providers. Most of the providers are global or local business actors. This challenges the implementation of the strategies. Legislations, the forming of agreements as well as the defining of requirements for products and systems are typical means for enhancing and controlling cyber security in society. This paper addresses the implementation of cyber security strategy from social systems and the leadership activities point of view. Especially, the paper studies leadership activities with the implementation of cyber security strategy in the public-private relations area. A preliminary leadership matrix model is presented for illustrating and classifying activities typically related to leadership activities in an organization: governance, leadership, managing and administration. A social system model is referred for increasing understanding about the influence of these activities on complex systems involving human activities. The leadership matrix model and social system model are applied in a case study for demonstrating the approach. The empirical data of the case study consist of a cyber strategy implementation plan. The paper classifies the activities of the implementation plan according to the leadership matrix model by using the content analysis research technique. This illustrates the focus of the implementation plan from the leadership activities point of view. These results are studied with the social system model for having a preliminary view of the influence of the implementation plan. This preliminary view is compared with the evaluations of the impacts of the activities of the implementation plan.

Keywords: leadership, cyber security strategy, system modelling, content analysis

1. Introduction

Global information networks host a variety of interdependent actors, activities and systems that adopt novel ways to implement activities on and by these networks. Nation states published cyber security strategies few years ago for setting national targets for security on global information networks. The reaching of the targets presented in the strategies often requires interaction and collaboration with the information and communication system and services providers. Most of the providers are global or local business actors. This means that the governmental and non-governmental organizations have to collaborate in all security situations for assuring cyber security and defending the vital functions of the society from threats. Especially, the leadership, decision-making, management and intelligence activities have to be shared across organizations for reaching the security targets both in the physical space and in cyberspace.

Legislations, the forming of agreements as well as the defining of requirements for products and systems are typical means for enhancing and controlling cyber security in public-private relations. This paper has a broad view on cyber security strategies and addresses them from social system's perspective and leadership activities point of view as depicted in Figure 1. Especially, the paper studies leadership activities with the implementation of cyber security strategy in the public-private relations area.

Leadership is a popular term typically defined as 'the office or position of a leader', 'capacity to lead', 'the power or ability to lead other people' or 'the act or an instance of leading' (Merriam-Webster 2014). One of the definitions of leading is 'providing direction or guidance' (Merriam-Webster 2014). The leadership models and methods are typically derived from philosophy, ethics, sociology and psychology, business economics as well as from case studies in organizations. Often these models and methods rise and fall on hype cycles. They are applied and improved but the understanding about their foundations still needs to be increased. Leadership is a practice without a comprehensive theory that would explain its semantics and principles.

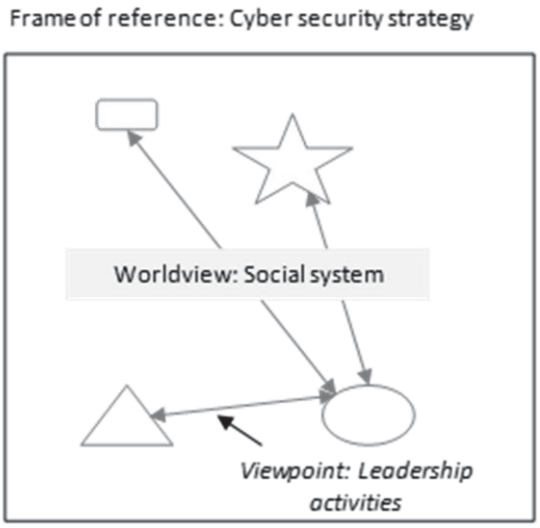


Figure 1: This paper addresses cyber security strategy from social systems and especially leadership activities point of views

The paper studies concepts and terms related to leadership. It presents a preliminary leadership matrix model for illustrating and classifying leadership activities in an organization: governance, leadership, managing and administration. The paper refers to a social system model for increasing understanding about the influence of these activities on complex systems involving human activities. The paper applies the leadership matrix model and social system model in a case study for demonstrating the approach. The empirical data of the case study consists of a cyber strategy implementation plan. The paper classifies the activities of the implementation plan according to the leadership matrix model by using the content analysis research technique. This illustrates the focus of the implementation plan from the leadership activities point of view. These results are studied with the social system model for having a preliminary view of the influence of the implementation plan. This preliminary view is compared with the evaluations of the impacts of the activities of the implementation plan.

2. The leadership matrix

Some concepts and terms related to leadership are studied next for increasing understanding about leadership activities. In an organization, leadership and decision-making are often related to governance, managing and administration. Governance provides a comprehensive view on organisation and its activities. It is often defined as 'the way that organizations or countries are managed at the highest level, and the systems for doing this' (Cambridge 2014) or 'the way that a city, company, etc., is controlled by the people who run it' (Merriam-Webster 2014). Governance includes creating and maintaining the high-level structure, architecture, guidelines, statutes and procedures of an organization. It supports leadership and enables managing and administration.

Managing is often defined as 'to take care of and make-decisions about' (Merriam-Webster 2014), 'to succeed in doing something, especially something difficult' or 'to be responsible for controlling or organizing someone or something, especially a business or employees' (Cambridge 2014). Administration is typically defined as 'the arrangements and tasks needed to control the operation of a plan or organization' (Cambridge 2014). Administration is supporting and assuring the organization to lead, manage and implement activities according to the regulations.

Managing is related to control, which Hamel (2012) regards managers to deify a lot. He argues it is time for a new ideology based on freedom and self-determination. Alberts et al. (2003) present similar thinking by including control free or self-syncronization in the classification of command and control approaches of information age. The other approaches of his classification are cyclic, interventionist, problem-solving, problem-bounding and selective control. Self-syncronization is the least control containing approach. Its principle is to reduce control and give more freedom for subordinates to perform their tasks.

The results of the analysis of the concepts of leadership, governance, managing and administration from the structure and action point of views are the following interpretations of these concepts:

- Leadership: As a structure a position of a leader and as an activity the ability to provide direction.

- Governance: As a structure contains the institutions, laws, and practices and as an activity overarching controlling.
- Managing: No structural dimension, as an activity the use of available means and resources for accomplishing an end.
- Administration: As a structure contains the body of persons who administer and as an activity to manage or supervise the execution.

A generalized and simplified model about the relations of leadership, governance, managing and administration are described as a leadership matrix in Figure 2. Leadership and managing are implemented with humans on social systems while governance and administration focus more on artefacts. Leadership and governance have a generalized view on people and things while administration and managing are more dealing with details.

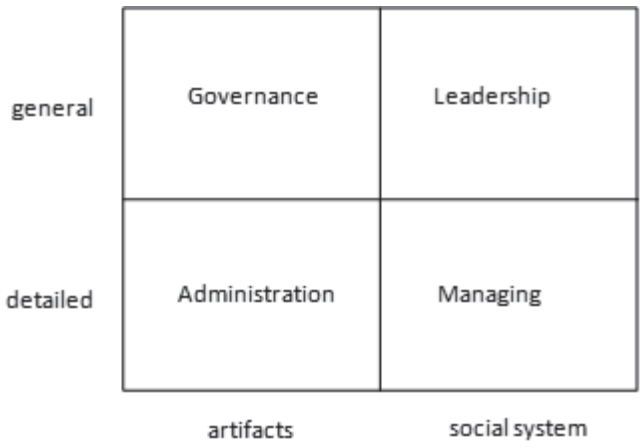


Figure 2: The leadership matrix model

A leader applies leadership, governance, managing and administration for setting and reaching strategic aims and planning and implementing operations with people and technology. If one of these activities is not concerned or if one of them is overemphasized the activities of a leader are not balanced. For example, heavily emphasized administration or bureaucracy is considered a burden that needs to be decreased. The aim of the leadership matrix model, however, is to support the visualizing of the volume of the activities. The target share of the activities depends on, e.g., mission, situation, people and technology.

3. The leadership matrix in relation to action levels

Activities are often classified to strategic, operational art and tactic level activities and operating. The relationships between these levels are described in Figure 3. A common definition for strategic is that it is 'of or relating to a general plan that is created to achieve a goal in war, politics, etc., usually over a long period of time' (Merriam-Webster, 2014). Strategic activities typically include the setting of the overall objectives for an organization and the determining of the path to these objectives or the developing of already chosen paths.

The target of operation art is to create such compositions and resources that enable success in military operations. It contains the creative application of knowledge, practice, cognition, imagination and intuition of a group of individuals. US DoD (2013) has the human capabilities perspective on operational art. It defines that 'operational art is the cognitive approach by commanders and staffs--supported by their skill, knowledge, experience, creativity, and judgment--to develop strategies, campaigns, and operations to organize and employ military forces by integrating ends, ways, and means'. A comprehensive view on operational art is to regard it as creating the right kind of resources and beneficial compositions to take successful steps towards the strategic objectives (Kuusisto et al. 2015). Cyberspace provides opportunities for the creating of novel compositions and dynamic resources. The utilization of these opportunities demands, however, understanding beyond spatially and temporally imminent events and already known means.

Tactic activities or tactics is often defined as 'an action or method that is planned and used to achieve a particular goal' (Merriam-Webster, 2014). Huttunen (2010) studies tactics and considers tactics as 'the optimal planning and applied use of resources and means arranged for mission accomplishment to achieve the goals in a combat. Tactics require knowledge about the means to combat and apply these means in practice'. Operating is the

activity phase where the activities are put in practice. Operating is typically defined as 'relating to the way a machine, vehicle, device, etc., functions or is used and controlled' (Merriam-Webster, 2014).

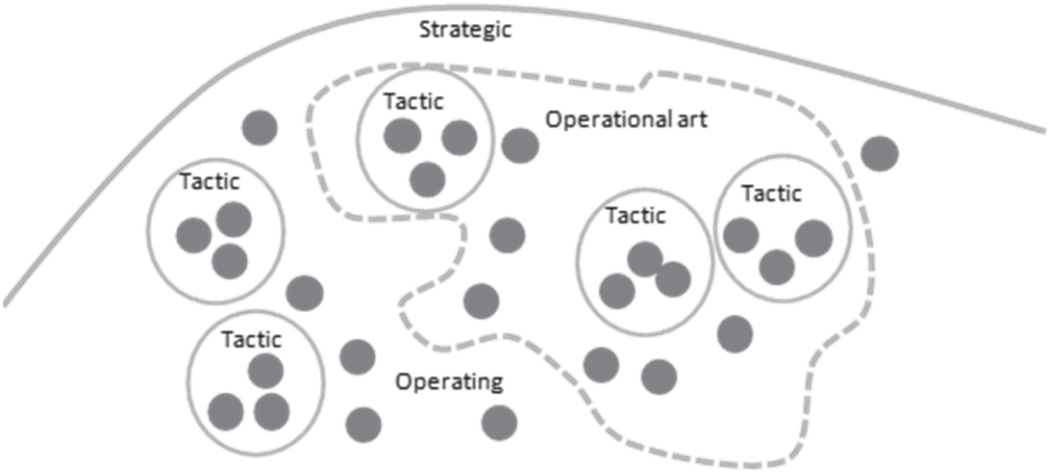


Figure 3: Operational art activities are challenged by strategic aims and tactic and operating level opportunities and restrictions, modified (Lund et al. 2005)

Tactic activities or tactics is often defined as 'an action or method that is planned and used to achieve a particular goal' (Merriam-Webster, 2014). Huttunen (2010) studies tactics and considers tactics as 'the optimal planning and applied use of resources and means arranged for mission accomplishment to achieve the goals in a combat. Tactics require knowledge about the means to combat and apply these means in practice'. Operating is the activity phase where the activities are put in practice. Operating is typically defined as 'relating to the way a machine, vehicle, device, etc., functions or is used and controlled' (Merriam-Webster, 2014).

Leadership, governance, managing and administration are implemented at all activity levels as outlined in Figure 4. Governance and administration are often emphasized at strategic level while operational art and tactics are often related to managing. Leadership approaches such as deep leadership are most often considered as operating level means.

Figure 5 (Kuusisto & Kuusisto 2006) outlines high-level information flows between the structural organizational levels. The discussions at the highest strategic structural level such as at the international organization level or at a state level address questions: why an organization or a state exists, what it should do and how its activities should be implemented? When perceived at the next lower level the descriptions of what the higher-level organization should do give the reasons for the lower-level organization to exist. The descriptions of how the higher-level organization should implement its activities give the basis for the descriptions of what the lower-level organization should do. For example, ministry of defence is assigned a task to form military forces with competent resources for defence. This task gives the reason for national defence university to exist.

Leadership addresses all the three questions why, what and how at all structural organizational levels. Governance typically produces descriptions for the questions of what and how. Managing and administration activities are mostly focused on how things are implemented. Leadership at strategic level gives people of an organization or its unit reasons to exist and directions to proceed. Governance at strategic level sets overall structure, guidelines and procedures for an organization or its unit. It gives basis for managing and administration at strategic level and governance at the operational art level. Governance at operational art, tactic and operating levels guides managing and administration at the level in concern.

Governance at a lower-level is often regarded as managing and administration when perceived at the higher-level. This is visualized in Figure 5 where higher-level how is related to lower-level what, i.e., the higher-level managing and administration are considered as governance at the lower-level.

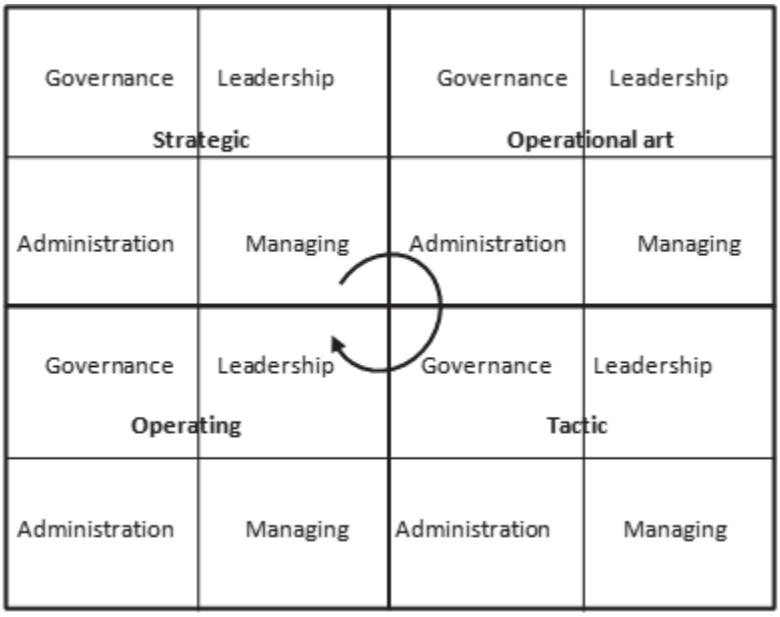


Figure 4: Governance, leadership, managing and administration are implemented at strategic, operational art, tactic and operating levels

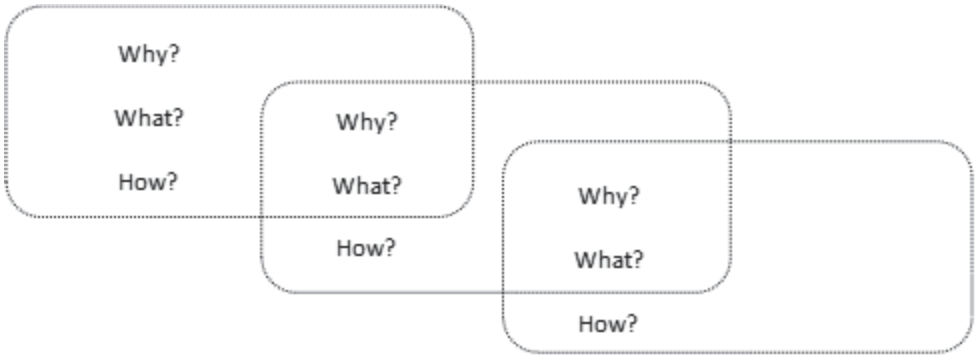


Figure 5: The higher-level organization's descriptions to the questions how and what give basis for the lower-level organization descriptions to the questions why and what (Kuusisto & Kuusisto 2006)

Leadership at operational art level is typically considered to cover abilities to perform operational art activities with people. In addition, leadership gives people reasons to exist and directions to proceed. At operational art level this means inspiring and committing people to create compositions and resources and giving directions for selecting steps on the strategic path.

Governance at operational art level contains structures and guidelines for selecting steps on the path towards strategic aims by creating compositions and resources, i.e., structure and guidelines of people and technology. Managing at operational art level includes decision-making about operational art issues. Administration at operational art level covers the support for implementing operational art activities according to regulations.

4. Leadership in social systems

Kuusisto (2004) presents a social system model derived from complexity thinking, system modelling, communication and cognition philosophy and sociology. The model is outlined in Figure 6 (Kuusisto & Kuusisto 2014). It consists of information, structure and action layers. The entities of the layers and information flows connecting these entities are derived from Aristotle's, Bergson's (1911), Parsons' (1951) and Habermas' (1984, 1989) thinking.

The information flows are shaping the social system. They are depicted with arrows in Figure 6. Information is flowing from values to norms through culture and pattern maintenance and to goals through community and integration and to polity, goal attainment, facts of present, organization, adaptation and finally back to values.

In addition, information flows from action to its neighbouring information class and external information enters from facts of present. The external world is influenced by the social system model's actions driven by goal attainment. The model is thus complex and emergent.

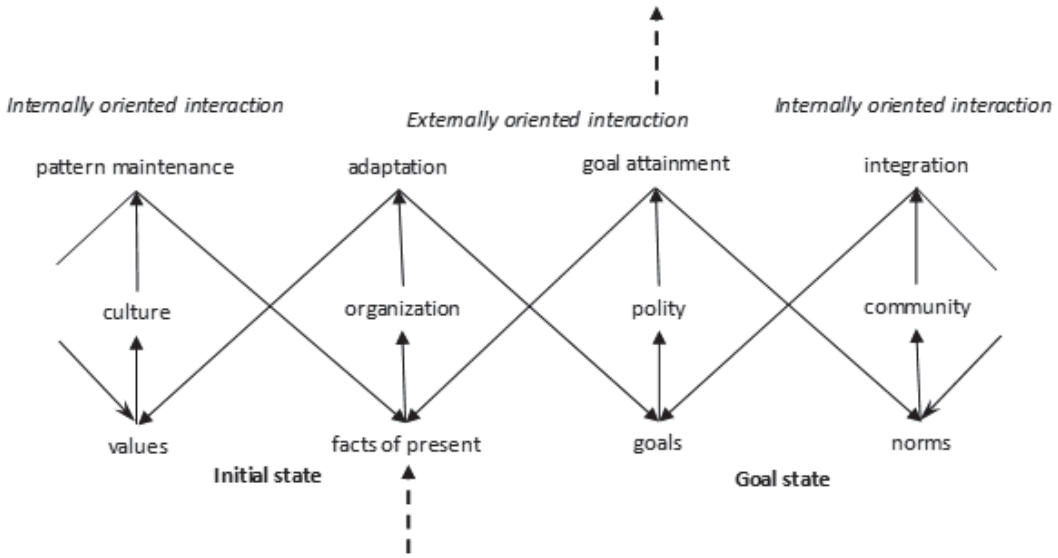


Figure 6: The social system model (Kuusisto & Kuusisto 2014)

The activities of the social system model are adopted from Parsons' (1951) outline of a social system (AGIL). They are adaptation (A), goal attainment (G), integration (I), and pattern maintenance or latency (L). Pattern maintenance is maintaining the stability of culture and its values through the processes which articulate values with belief systems such as ideology (Parsons 1985). In cyber era it includes maintaining patterns defining the structure of digital society such as identity and privacy. Leadership in social systems is outlined by studying the concepts of the leadership matrix, i.e., leadership, governance, managing and administration and the concepts of the social system model. The result is the coverage of leadership, governance, managing and administration in the social system model presented in Figure 7. As figure 7 shows the concepts of the leadership matrix are overlapping in the social system. Governance is related to all actions as well as culture, organization and community. Leadership emphasizes goal attainment, polity and integration. Administration is most close to adaptation and organization. Managing is mostly related to adaptation and goal attainment. Figure 8 visualizes the activities of the leadership matrix that need to be addressed when aiming to influence on the actions and structure of a social system. Leadership should be strengthen when the goal attainment or integration activities or polity structure of a social system need to be improved. Governance is an enabler of all the actions and culture, organization and community development. Especially, if pattern maintenance, culture or community is challenged then there is a need to stress governance. Managing can be applied to improve adaptation to the current situation or goal attainment. Administration mostly strengthens adaptation to the current situation and organization structure.

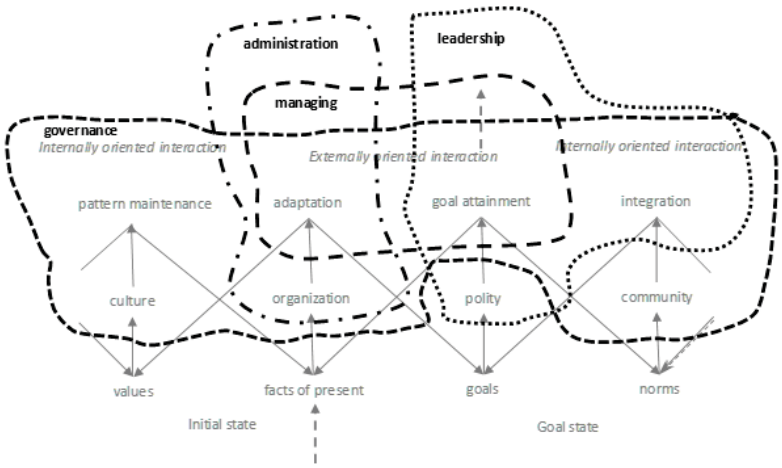


Figure 7: The coverage of leadership, governance, managing and administration in the social system model

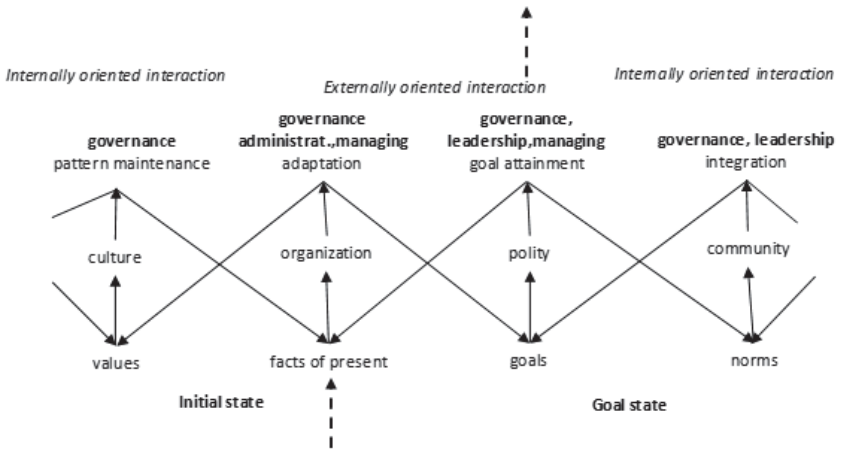


Figure 8: Activities of the leadership matrix that need to be addressed when aiming to influence on the actions and structure of a social system

For example, a cyber strategy implementation plan that mainly consists of governance, administration and managing activities is most likely to influence on the adaptation to the current situation. Leadership activities needs to be increased and administration activities decreased for making change and influence on goal attainment and integration.

5. A case study

The leadership matrix model and social system model are applied in a case study for demonstrating the proposed approach. The empirical data of the case study consist of the implementation plan of Finland's Cyber Security Strategy. Finland published Cyber Security Strategy as a Government Resolution in 2013 (Finnish Government 2013). It defines vision and the key objectives for protecting society and its vital functions against cyber threats. The aim of the strategy is that Finnish society will gain the benefits of digitalization in a secured way. Finland as a country with a population of only about 5 million people has a tradition for a strong collaboration and support between authorities. The Security Committee (2016) is a formal body assisting the Government and ministries for enhancing comprehensive security and coordinating proactive preparedness. The Security Committee guided the forming of Cyber Strategy and decisions to apply the collaboration model between authorities to cyber world related activities.

The implementation plan of the Cyber Security Strategy was published in 2014 (The Security Committee 2014). It consists of 74 activities. These activities include 56 activities that are related to the public-private relations area. So, the collaboration models between the authorities and actors applied in Finland needed to be extended for cyber world to include private companies.

The activities of the implementation plan were classified according to the leadership matrix model by following Krippendorff's (2013) content analysis research technique. Content analysis is a growing research technique 'for making replicable and valid interferences from texts (or other meaningful matter) to the contexts of their use' (Krippendorff 2013). Each activity of the strategy implementation plan was placed on all the categories it is related to. The results of the analysis are shown in Figure 9. The results of the analysis of the activities related to the public- private relations area are in parenthesis.

The clear focus of the activities is on managing and administration as depicted in Figure 9. One third of the activities are considered as governance or leadership. As shown in Figure 7 and 8, managing and administration mostly have an influence on adaptation, organization and goal attainment. So, when the results are studied with the social system model it can be assumed that the implementation plan would support the adaptation to the current competences, processes, technology and organization structure as well as goal attainment.

The share of administration activities decreases and the share of leadership activities increases when the results of the analysis of all the activities are compared to the analysis of the activities related to the public- private area. The administration activities of the strategy implementation plan are internally focused on government while the public-private area related activities emphasize managing and leadership. When the analysis of the

activities related to the public-private area are studied with the social system model it can be assumed that the activities related to the public-private area would support goal attainment and have an effect on polity structure.

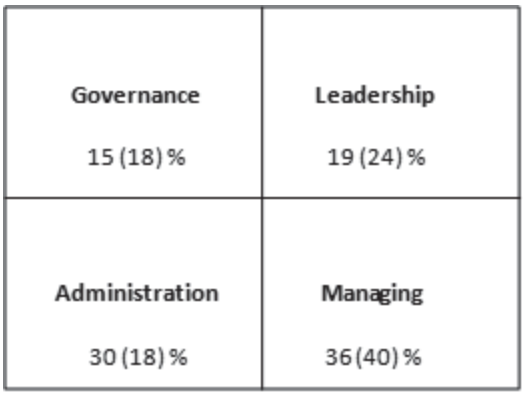


Figure 9: The classification of activities of a cyber strategy implementation plan according to the leadership matrix model. The classifications of the activities related to the public-private relations area are in parenthesis

In 2015 the status and impacts of the strategy implementation plan were evaluated by the government. The activities that were considered to have had most impact are:

- The design, implementation, delivery and use of the Government Security Network (TUVE) and sector-independent ICT services (TORI) produced by Government ICT Centre Valtori established in 2014.
- The establishment of the National Cyber Security Centre Finland (NCSC-FI). It collects and delivers information on the cyber security situation in Finland and supports actors for protecting and recovering from cyber threats (NCSC-FI 2016).
- The providing of a special training course for authorities about cyber security.
- The establishment of SecICT body for the central government information security management. SecICT is a network of experts managing information and cyber related incidents affecting the central government.
- The establishment of JYVSECTEC cyber security research, development and training centre in Jyväskylä in Finland (JYVSECTEC 2016).

The major characteristic of these activities is that they have created novel organizational or decision-making structures or changed the existing structures. In addition, they have improved performance and competence by enhancing processes or delivering new ICT services or training.

In addition to these activities that were considered to have had a significant impact there were together 16 activities that were considered to have had significant or clear impact on cyber security of society. Most of these 16 activities are related to the public-private relations area. Two third of these 16 activities were classified as governance or managing activities. Especially, the share of governance activities increases compared to the classification results of all activities. So, even if the share of the governance activities is quite low in the implementation plan, half of them were considered as activities having significant or clear impact.

As shown in Figures 7 and 8, governance and managing influence on adaptation and goal attainment. In addition, there is a potential to influence on pattern maintenance, integration, culture, organization and community. The preliminary view of the impacts of the strategy implementation plan emphasised adaptation and goal attainment. The analysis of the activities of the implementation plan considered having significant or clear impact indicates that in addition to adaptation and goal attainment the plan influences on pattern maintenance, integration, culture, organization and community. This is visualized by studying the activities of the implementation plan having significant or clear impact with the leadership matrix and social system model.

6. Conclusions

This paper proposes a theoretically motivated approach and demonstrates it with a small case study about cyber strategy implementation plan. The approach contains a leadership matrix model and a social system model. The leadership matrix model consists of governance, leadership, managing and administration. The paper refers to a

social system model for increasing understanding about the influence of these activities on complex systems involving human activities.

The research preliminary verifies the approach by conducting a case study. It shows that the proposed approach is plausible. More empirical studies are needed to continue the validation of the approach. The case study, however, indicates that the proposed approach increases understanding about the influence of activities on social systems. The empirical data of the case study consist of Finland's Cyber Security Strategy implementation plan published in 2014. The paper shows that the clear focus of the activities of the plan is on managing and administration. One third of the activities are considered as governance or leadership. When these results are studied with the social system model it can be assumed that the implementation plan would support the adaptation to the current competences, processes, technology and organization structure as well as goal attainment. The impacts of the plan were evaluated by the government in 2015. The study of the evaluation results shows that the majority of the activities considered to have had a significant or clear impact were governance or managing activities having influence on adaptation and goal attainment as well as on pattern maintenance, integration, culture, organization and community. This is visualized by studying the activities of the implementation plan having significant or clear impact with the leadership matrix and social system model.

References

Alberts, D.S., Hayes, R. E. (2003). Power to the Edge: Command and Control in the Information Age. CCRP, USA. 283 s.

Bergson, H. (1911). *Creative Evolution*. University Press of America.

Cambridge (2014). *British English Dictionary and Thesaurus, Cambridge Dictionaries Online*. Viewed 16 December 2014, http://dictionary.cambridge.org/dictionary/british/

Finnish Government (2013). *Finland's Cyber Security Strategy*. Government Resolution 24.1.2013. Retrieved on the 2nd of April 2016 from http://www.turvallisuuskomitea.fi/index.php/en/component/k2/38-cyber-security-strategy

Habermas, J. (1984). *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Boston, MA: Beacon Press.

Habermas, J. (1989). *The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason*. Boston, MA: Beacon Press.

Hamel, G. (2012). *What Matters Now: How to Win in a World of Relentless Change*. Ferocious Competition, and Unstoppable Innovation. Jossey-Bass, San Francisco, USA.

Huttunen, M. (2010). Monimutkainen taktiikka. Taktiikan laitoksen julkaisusarja 1, n:o 2/2010, Edita Prima Oy, Helsinki, 2010. In Finnish.

JYVSECTEC (2016). Cyber Security Solutions through Research, Development and Training. Retrieved on the 2nd of April 2016 from http://jyvsectec.fi/en/

Krippendorff, K. (2013). *Content analysis: an introduction to its methodology*, 3rd edition. Sage, Newbury Park, CA, USA

Kuusisto, R. (2004). *Aspects on availability*. Edita Prima Oy, Helsinki, Finland.

Kuusisto, T., Kuusisto, R. (2006). *Verkostopuolustuksen johtaminen – tietovirtojen näkökulma itsesynkronoitumiseen*. Tiede ja ase, Suomen Sotatieteellisen Seuran vuosijulkaisu N.o 64, 2006, pp. 37-56. In Finnish.

Kuusisto, T., Kuusisto, R. (2014). *Prerequisites for Creating Resources and Compositions for Cyber Defence*. Proc. of 15th Australian Conference on Information Warfare, 2014 SRI Security Congress, Perth, Australia, 1.-3.12.2014, 8p.

Kuusisto, T., Kuusisto, R., Roehrig, W. (2015). *Situation Understanding for Operational art in Cyber Operations*. In Abouzakhar, N. (ed.) Proc of the 14th European Conference on Cyber Warfare and Security ECCWS-2015, Hatfield, UK, 2.-3.7.2015, pp. 169-178

Lund, O.-P., Mikkola, T., Kuusisto, R., Kuusisto, T. (2005). *Military Activities on Organizational Levels – Information Needs for High-Level Tactical Decision-Making*. In Kuusisto, R. & Rantapelkonen, J. (eds.) Struggling to understand information war, National Defence College, Department of Leadership and Management Studies & Deparment of Tactics and Operations Art, Publication Series 2, Article Collection, Finland, 2005, pp. 134-148

Merriam-Webster (2014). *Merriam-Webster online dictionary*. Retrieved on 25 March 2014 from http://www.merriam-webster.com

NCSC-FI (2016). *The National Cyber Security Centre Finland*, retrieved on the 2nd of April 2016 from https://www.viestintavirasto.fi/en/cybersecurity.html

Parsons, T. (1951). *The Social System*. Free Press, Glencoe, IL.

The Security Committee (2014). *The implementation programme of the Cyber Security Strategy.* A summary in English, the original version in Finnish, Finland, 11 March 2014. The English summary retrieved on 19 February 2016 from http://www.turvallisuuskomitea.fi/index.php/en/component/k2/39-the-implementation-programme-of-the-cyber-security-strategy. The original version retrieved on 19 February 2016 from http://www.turvallisuuskomitea.fi/index.php/fi/kyberturvallisuusstrategia

The Security Committee (2016). *What is the Security Committe*e? Retrieved on the 2nd of April 2016 from http://www.turvallisuuskomitea.fi/index.php/en/

US DoD (2013). *JP1-02, Dictionary of Military and Associated Terms 2010*, amended 2013. Retrieved on 13 May 2014 from http://www.dtic.mil/doctrine/dod_dictionary/

Cyber Security Capability and the Case of Finland

Martti Lehto[1] and Jarno Limnéll[2]
[1]University of Jyväskylä, Finland
[2]Aalto University, Finland
martti.lehto@jyu.fi
jarno.limnell@aalto.fi

Abstract: Many countries are building their national cyber security capabilities. They are defining what they mean by cyber security in their national strategy documents. The common theme from all of these varying definitions, however, is that cyber security is fundamental to both protecting government secrets and enabling national defense, in addition to protecting the critical infrastructure that permeate and drive the 21st century global economy. The development of cyber security capabilities is a complex matter. Whether at the nation state level, or in an enterprise, various factors need to be taken into consideration. A layered approach can provide more comprehensive coverage than single, disparate solutions. The measurements of security postures and progress over time are important elements to strengthening policies, evaluating risks and anticipating future scenarios. Different cybersecurity indices have been published in the past few years, yet not all measure the same capabilities. According to International Telecommunication Union (ITU) Cyber security is not an end unto itself; cyber security must be understood as a means to an end. The goal should be to build confidence and trust that critical information infrastructure would work reliably and continue to support national interests even when under attack. Therefore the focus of national cyber security strategies should be on the threats most likely to disrupt vital functions of society. Digitalization and information societies are ever evolving and new cyber threats continue to be devised. In this progress, cyber security must form an integral and indivisible part of the nation's security process. Countries need to be aware of their current capability level in cyber security and at the same time identify areas where cybersecurity needs to be enhanced. It can be said that cyber security is a constant "arms race" between countries, but also between the security community and the hostile hackers. Today's high-profile cybersecurity incidents have underlined the crucial importance of strengthening cyber resilience in general, as well as the protection of critical infrastructure from cyber threats in all countries. In order to achieve these goals, public and private stakeholders need to be equipped with the capacity to effectively prevent, mitigate and respond to cyber-attacks and incidents. Resilience stands for the continuation of operations even when society faces a severe disturbance in its security environment, the capability to recover quickly from the shock, and the ability to either remount the temporarily halted functions or re-engineer them. How should one measure the maturity of national cyber security? Agreed international standards which can measure the level of national cyber capability, are not available. While nations are developing the cyber capabilities to operate in the cyber realm, measuring national cyber capabilities remains problematic. However, several different organizations have made their own analyses of the national cyber security capabilities, but especially measuring military cyber capabilities is challenging because of information classification. Countries prefer to be secretive about their military cyber capabilities. Another challenge in measuring national cyber capabilities relates to the ubiquity and dual-use nature of computing and cyber tools, the stealth and immediacy of cyber operations and uncertainty over the responsibilities of civilian and military organizations (International Institute for Strategic Studies 2014). In this paper we use DOTMLPF-II as a research framework. We analyze the cyber security capability using the DOTMLPF-II components: Doctrine, Organization, Training, Materiel, Leadership, Personnel, Facilities, Interoperability and Information. DOTMLPF-II analysis is the first step in building the national cyber security capability building. It determines necessary recommendations which are required to fill a capability gap identified in the analysis. As empirical material we use the reports and studies of European Union, BSA Software Alliance, Global Cybersecurity Index of the International Telecommunication Union and ABI Research, and Microsoft Intelligence report. On the basis of DOTMLPF-II cyber security capability modelling we analyze the national cyber security capability in Finland based on the reports mentioned above.

Keywords: cyber security, capability, DOTMLPF-II

1. Introduction

Cyber security capability is the new element of a nation's defense and security policy. Cyber warfare is one element of the hybrid warfare which blends conventional warfare, irregular warfare and cyber warfare. Referring to war, the cyber instrument has become a domain like land, sea, air and space. Nation-states are spending more money in order to create their cyber capabilities and the role of using cyber domain has become emphasized in National security and Military strategies. The development can be seen as the beginning of the "digital arms race", where the rules of engagement are not yet codified. Just five years ago, it was considered by many to be science fiction that nation states could use bits and bytes to create the same sort of destruction as bullets and bombs. Now this is becoming reality.

Understanding (and estimating) cyber capabilities requires – as a starting point – analysis of states´ strategic, technological and political intentions. Available doctrines reveal something about the capabilities but often

resource allocation, such as financial and organizational investments to cyber capabilities, are perhaps the clearest indicator of state cyber warfare activity. Howeverthe financial figures are often vague or kept in secret. Many states have also announced the formation of cyber units in their armed forces which can be measured as part of creating cyber capabilities. Other indicators of cyber activity are often said to be the recruitment of cyber experts, adapting or upgrading military cyber strategies and the level of sophistication of states´ public-private-partnership. Analysis involves understanding how states themselves view the cyber domain since there are varying conceptions of the terms "cyber" and "cyber warfare." For example, what might be considered in the West to be separate concepts (like cyber warfare and information warfare) is more blended elsewhere.

Cyber security has quickly evolved from a technical discipline to a strategic concept. Since cyber security has evolved from a technical discipline to a strategic concept, and because cyber attacks can affect national security at the strategic level, cyber security must look beyond the tactical arena. Strategic planning, as a structured and systematic process, is successful when it is leader-led and consists of effective tools for capacity building. The strategic planning process is where leaders of an organization establish the vision of the organization's future and then develop and implement the actions necessary to achieve that future. Holistic approach is needed in both analysis and development of cyber capabilities.

One example of a holistic approach to possible indications of cyber capability is presented in the International Institute for Strategic Studies´ book "The Military Balance". The table 1 (IISS 2014) describes well the wide range of indicators which should be taken into consideration when assessing a nation´s cyber warfare capabilities. However, it is an open question on how to integrate and measure these indicators to overall assessment of nation´s cyber capability.

Table 1: Indicators of the nation's cyber security capability

Political	Military	Economic	Social	Information/ Technology	Infrastructure	Other
Political system	Military cyber startegy and doctrine	Defence dudget, service budgets	Maturity of Information society	High-tech density	Military netwroks	Manufacturers´ advertisements
Social stability	Organisational structure/units	Program budgets	Top technical Universities	Know-how, experience	SIGINT + platforms	Strategic purchases and sales
National ambitions	Education, training, exercises	GDP	Post-graduate students	Innovation	Communications	Unusual policy/security attention
International standing	Known/ suspected operations	Raw materials	Graduation in science, engineering	State-of-the-art technology	High-speed access	
Relationship with hackers	Intelligence and fusion	Export/import restrictions	Known hackers	Retail electronics	Advanced services	
Regulatory action	Materiel, logistics, infrastructure	Acquisitions, procurements	R&D intensity	Advanced technologies	Number of ISPs	
Parliamentary discussions, security documentations		Patents, R&D funding	Researcher concentration		Space exploration capabilities	
		Hich-tech public companies			Industrial base	
		Products with high R&D loads				
		Manufacturing capability				

Finland is one of the most developed information societies whose functioning relies on various electronic networks and services. The level of digital dependence of the society should also be noted since it directly affects how devastating successful cyber attacks may be. Finland is deeply cyber dependent in every societal aspect – that is – as a state, as an economy comprising financial activities of enterprises, and as a society trying to maintain its citizens' current way of life. Finland published its first National Cyber Strategy in 2013. The Strategy states that "Cyber security means the desired end state in which the cyber domain is reliable and in which its functioning is ensured." According to the Strategy, cyber security is built on sufficient capabilities development over the long term, their well-timed and flexible use and the resilience of society's vital functions against disturbances in cyber security. Finland´s Cyber Strategy also states that the government is responsible for

providing political guidance and strategic guidelines for cyber security as well as for taking the required decisions regarding the resources and prerequisites to be allocated to it.

In the Finnish model for a cyber-secured society, the role of the government has been given considerable emphasis to guarantee the development of a favorable atmosphere through infrastructure, legislation, and accessibility for all. In a holistic context Finland has a long tradition of public-private partnerships and a comprehensive approach to security is being also applied to cyber-security. The arrangements of comprehensive security are defined in a Government resolution which defines the principles of ensuring the vital functions, such as the population´s income security and capacity to function as a society.

2. Capability building model

DOTMLPF is a tool for Joint Force planning in US Armed forces. It has also an expanding version like DOTMLPF-P (Policy), DOTMLPF – FREE (F: Finances, R: Relationships, E: Efficiency, E: Effectiveness). In this research we use DOTMLPF-II model (I: Information, I: Interoperability). (Department of Defense 2015)

DOTMLPF-II stands for:

Doctrine: The doctrine analysis examines the way the military fights its conflicts with emphasizes on maneuver warfare and combined air-ground campaigns to see if there is a better way that might solve a capability gap.

Organization: The organization analysis examines how we are organizing to fight.

Training: The training analysis examines how we prepare our forces to fight tactically from basic training, advanced individual training, various types of unit training, joint exercises, and other ways to see if improvement can be made to offset capability gaps.

Materiel: The materiel analysis examines all the necessary equipment and systems that are needed by our forces to fight and operate effectively and if new systems are needed to fill a capability gap.

Leadership and Education: The leadership and education analysis examines how we prepare our leaders to lead the fight from squad leader to four-star general/admiral and their overall professional development.

Personnel: The personnel analysis examines availability of qualified people for peacetime, wartime, and various contingency operations to support a capability gap by restructuring.

Facilities: The facilities analysis examines military property, installations and industrial facilities (e.g. government owned ammunition production facilities) that support our forces to see if they can be used to fill in a capability gap.

Interoperability: The ability of the systems or troops to offer support and to receive it from other systems or troops so that the cooperation will be efficient.

Information: Demands of the data, information and knowledge which are needed in the capabilities and in the processes which have been designed to collect and to deal with.

3. Maturity of national cyber security

In this research we have used four international research reports which are focused to the nation's cyber security capability.

3.1 Global cybersecurity index, GCI

A joint collaborative project by the International Telecommunication Union (ITU) and Allied Business Intelligence (ABI) Research has created the Global Cybersecurity Index (GCI) which aims to measure and rank each nation state's level of cybersecurity development. (ABI 2015)

Following the ITU´s Global Cybersecurity Agenda (GCA) Framework, the GCI identifies 5 indicators

- 1. Legal
- 2. Technical
- 3. Organizational
- 4. Capacity Building
- 5. Cooperation

Rooted in the ITU Global Cybersecurity Agenda, the GCI looks at the level of commitment in five areas. The result is a country-level index and a global ranking on cybersecurity readiness.

Legislation is a critical measure for providing a harmonized framework for entities to align themselves to a common regulatory basis, whether on the matter of prohibition of specified criminal conduct or minimum regulatory requirements. Legal measures also allow a nation state to set down the basic response mechanisms to breach: through investigation and prosecution of crimes and the imposition of sanctions for non-compliance or breach of law. (ABI 2015)

Technology is the first line of defense against cyber threats and malicious online agents. Without adequate technical measures and the capabilities to detect and respond to cyberattacks, nation states and their respective entities remain vulnerable to cyber threats. Technical measures can be measured based on the existence and number of technical institutions and frameworks dealing with cybersecurity endorsed or created by the nation state. (ABI 2015)

Organization and procedural measures are necessary for the proper implementation of any type of national initiative. A broad strategic objective needs to be set by the nation state, with a comprehensive plan in implementation, delivery and measurement. The organizational structures can be measured based on the existence and number of institutions and strategies organizing cybersecurity development at the national level. (ABI 2015)

Capacity building is intrinsic to the first three measures (legal, technical and organizational). Understanding the technology, the risk and the implications can help to develop better legislation, better policies and strategies, and better organization as to the various roles and responsibilities. Capacity building can be measured based on the existence and number of research and development, education and training programs, and certified professionals and public sector agencies. (ABI 2015)

Cybersecurity requires input from all sectors and disciplines and for this reason needs to be tackled from a multi-stakeholder approach. **Cooperation** enhances dialogue and coordination, enabling the creation of a more comprehensive cybersecurity field of application. National and international cooperation can be measured based on the existence and number of partnerships, cooperative frameworks and information sharing networks. (ABI 2015)

The GCI does not seek to determine the efficacy or success of a particular measure, but simply the existence of national structures in place to implement and promote cybersecurity. Many countries share the same ranking which indicates that they have the same level of readiness. The index has a low level of granularity since it aims at capturing the cybersecurity commitment/preparedness of a country and not its detailed capabilities or possible vulnerabilities. (ABI 2015)

In the GCI United States of America is rank 1 by index 0.824. Finland is rank 8 by index 0.6176. Among the European countries Finland is rank 5 and the results are:

- Legal 0.5000
- Technical 0.6667
- Organizational 0.8750
- Capacity building 0.5000
- Cooperation 0.5000
- Total index 0.6176

The GCI analyses Finland's capability in the five areas. According to the GCI specific **legislation** on cybercrime has been enacted through the Criminal Code and Act on the Protection of Privacy in Electronic Communications. (ABI 2015)

Regarding the **technical measures** Finland has an officially recognized national CIRT (CERT-FI) and from 1.1.2014 The National Cyber Security Centre Finland (NCSC-FI). Finland's National Security Auditing Criteria the main goal of which is to harmonize official measures when an authority conducts an audit in a company or in another organization to verify their security level. The Accreditation of information security inspection bodies in Finland is regulated in the act on information security inspection bodies. (ABI 2015)

Regarding the **organization measures** Finland has an officially recognized national cybersecurity strategy since 2013 (Finland cybersecurity strategy). It defines the key goals and guidelines which are used in responding to the threats against the cyber domain and which ensure its functioning. By following the Cyber Security Strategy's guidelines and the measures required, Finland can manage deliberate or inadvertent disturbances in the cyber domain as well as respond to and recover from them. The Information Society Program provides a national governance roadmap for cybersecurity in Finland. The Security Committee monitors and coordinates the implementation of a national cybersecurity strategy, policy and roadmap by respective agencies. The National Emergency Supply Agency is involved in organizing sector-specific preparedness exercises on some critical infrastructure sectors. (ABI 2015)

Regarding the **capacity building** Finland does not have any officially recognized national or sector-specific research and development (R&D) programs/projects for cybersecurity standards, best practices and guidelines to be applied in either the private or the public sector. (ABI 2015) The research is based on information from 2013 and early 2014. Now the Tekes (The Finnish Funding Agency for Technology and Innovation) has funded the national Cyber Trust R&D Program (2015–2018), which will create new high-level competences in areas expected to be important for cyber education and businesses in the future. The projects have innovative visions and multidisciplinary approaches. The aim of these projects is to create a foundation for research and the industry to address the needs emerging in the cyber security domain.

Cyber security education in Finland is supported by the modern data network laboratory which focuses on the use of the information and communications technology (ICT) education as well as projects. According to the ABI research Finland does not have the exact number of public sector professionals certified under internationally recognized certification programs in cybersecurity. Finland does not have any certified government and public sector agencies certified under internationally recognized standards in cybersecurity. (ABI 2015) In October 2014 The Cyber security centre of the Finnish Communication Regulatory Authority has given to the KPMG IT Sertifiointi Oy right to operate as the cyber security evaluation institution.

To **facilitate sharing** of cybersecurity assets across borders or with other nation states, Finland has officially recognized partnerships with the following organizations: ITU, FIRST, European Government Certs group. Finland has officially recognized national or sector-specific programs for sharing cybersecurity assets within the public sector through its national CIRT/NCSC-FI. Finland does not have any officially recognized national or sector-specific programs for sharing cybersecurity assets within the public and private sector. (ABI 2015) The Ministry of Foreign Affairs in Finland has also established a position of cyber ambassador since 2014 and the cyber ambassador´s main task is to strengthen Finland´s national cyber security through active and efficient participation in the activities of international organizations and collaborative fora that are critical to cyber security.

3.2 EU cybersecurity dashboard

With an increasing focus on improving cyber resilience in both the Member States and at the EU level, the Business Software Alliance (BSA) report 2015 provides a comprehensive overview of the state of the current cybersecurity frameworks and capabilities. The report examines five key areas of each EU Member State's cybersecurity policy environment:

- Legal foundations for cybersecurity
- Operational capabilities
- Public-private partnerships

- Sector-specific cybersecurity plans
- Education

European Union Cybersecurity Maturity Dashboard (2015) is based on an assessment of twenty five criteria across five themes mentioned above.

In the area of **legal foundations** for cybersecurity Finland has Cyber Security Strategy and the Government Decision on the Security of Supply (2008). It is the latest set of official goals and standards relating to the protection of critical infrastructure. While there is no legislation or policy in place in Finland that requires the establishment of a written information security plan, the Ministry of Finance established Government Information Security Management Board (VAHTI) and published the Government Information Security Guideline in 2009. The Government Decree on Information Security in Central Government 2010 requires classification levels to be applied to information, which is a reflection of both the risk level involved in disclosing the information and the necessary security requirements to be complied when handling the information. There is no legislation or policy in place in Finland that requires each agency to have a chief information officer or chief security officer or requires mandatory reporting of cybersecurity incidents. Finland has developed some specific security requirements for procurement related to critical infrastructure or the handling of classified information (KATAKRI or National Security Auditing Criteria). (BSA 2015)

In the area of **operational capabilities,** the National Cyber Security Centre Finland (NCSC-FI) was established in 2014 through merging of CERT-FI and NCSA-FI. This body is responsible for the coordination of incident response, incident management, security incident reporting and information security measures for both government institutions and the private sector. NCSC-FI acts as the national competent authority for network and information security. Finland carries out national cyber exercises every half year. Finland has taken part in multi-national exercises organized by the European Union and NATO. (BSA 2015)

In the area of **public-private partnerships** the National Emergency Supply organization (NESO) is a network of multiple public-private partnership initiatives whose objectives are related to the security of supply. NESO is responsible for measures related to developing and maintaining the security of supply. The Finnish Information Security Cluster (FISC) is an association of Finnish information security companies. Their role is primarily business advocacy, however in representing the information security sector, FISC is significantly engaged with Finnish cybersecurity. (BSA 2015)

In the area of **sector-specific cybersecurity plans** the Government Decision on the Security of Supply 2008 addresses, in part, cybersecurity as it relates to sector specific critical infrastructure. Finland, however, does not have sector-specific joint public-private plans or security priorities in place. (BSA 2015)

Finland's Cyber Security Strategy 2013 includes a detailed commitment to cybersecurity **education**. It states that: "the study of basic cyber security skills must be included at all levels of education." (BSA 2015) (Finland's Cyber Security Strategy 2013)

There are no overall rankings or scores in this study. The results are shown as Yes/No/Partial results to the presented 25 questions. If we calculate together all "Yes" results so the total top results are:

- Austria 18
- Estonia 17
- Czech Republic, Finland and Germany 16

3.3 Microsoft security intelligence report

The Microsoft Security Intelligence Report (SIR) Worldwide Threat Assessment focuses on software vulnerabilities, software vulnerability exploits, malware, and unwanted software. The Nordic countries, including Denmark, Finland, Iceland, Norway, and Sweden, have perennially been among the healthiest locations in the world with regard to malware exposure, as has Japan. The infection and encounter rates for these locations were typically about half of the worldwide averages. (Microsoft 2015)

The locations with the most computers reporting as fully protected by real-time security software include Finland, with 83.9 percent of computers reporting as fully protected, Denmark at 79.5 percent, and Norway, at 78.9 percent. The ranking of countries and regions by unprotected rate is largely an inverse of their ranking according to protected rate. The locations with the fewest computers reporting as fully unprotected include Finland, at 10.4 percent, Denmark at 14.2 percent, and the Czech Republic at 14.4 percent. (Microsoft 2015)

Locations with large concentrations of malware hosting sites included Brazil (41.0 per 1,000 Internet hosts), Costa Rica (38.8), and Russia (23.9). Locations with low concentrations of malware hosting sites included Taiwan (2.8), Saudi Arabia (4.3), and Finland (4.4). (Microsoft 2015)

3.4 EU impact assessment

EU Commission Staff Working Document - Impact Assessment covers policy options to improve the security of the internet and other networks and information systems underpinning services which support the functioning of our society. (EC 2013)

Member States have very different levels of capabilities. This situation hinders the creation of trust among peers in the Member States which is an important prerequisite for cooperation and information sharing. According to a market study, Member States can be divided into four groups on the basis of the maturity of their Network and Information security (NIS) markets:

- Group 1, the Champions: Denmark, Finland, the Netherlands, Sweden, the United Kingdom
- Group 2, the Pillars: Austria, Belgium, Germany, Luxembourg, France, Ireland

These two groups together represent69% of the EU GDP but 82% of total security spending. These clusters are characterized by high average security spending, a strong presence of high profile security business users, and greater adoption of advanced security solutions. (EC 2013)

Group 3, the Runners Up include the Southern European countries: Cyprus, Greece, Italy, Malta, Portugal, and Spain and: Czech Republic, Hungary and Slovenia: this cluster shows some delay with the advanced clusters but a good potential for growth. (EC 2013)

Group 4, the Learners: Bulgaria, Estonia, Latvia, Lithuania, Poland, Romania, and Slovakia. This cluster includes the remaining Member States with the lowest level of NIS spending and maturity. (EC 2013)

The table 2 summarizes the information provided by the Member States on their national capabilities. According to the information received, only group 1 countries and a large majority of group 2 countries have a level of preparedness which corresponds to the targets pursued by the Commission since 2009 (CIIP Action plan and CIIP Communication of 2011). (EC 2013)

Table 2: National network and information security capabilities

Group of countries	N/G CERTs	CERTs EGC group	NIS Strategy	Contingency/Cooperation Plan
1 - DK, FI, NL, SE, UK	DK, FI, NL, SE, UK	DK, FI, NL, SE, UK	DK, FI, NL, SE, UK	DK, FI, NL, SE, UK
2 - AT, BE, DE, FR, IE, LU	AT, BE, DE, FR, IE, LU	AT, DE, FR,	AT, DE, FR, IE, LU	AT, DE, FR, LU
3 - CY, GR, IT, MT, PT, ES, CZ, HU, SL	CY, GR, IT, MT, PT, ES, CZ, HU, SL	ES, HU	CY, EL, ES,CZ, HU	CY, EL
4 - BG, EE, LV, LT, PL, RO, SK	BG, EE, LV, LT, PL, RO, SK		EE, LV, LT, PL, RO, SK	EE, LV

Not all Member States have an operational national/governmental CERT in place to handle NIS incidents and prevent them from happening by monitoring threats. Only some Member States have adopted national cyber security strategies. Not all Member States have in place a cyber-incident contingency/cooperation plan, providing protocols for communications and coordinated action in crisis situations, and not all Member States have carried out or regularly carry out cyber incident exercises, which are major tools to put in place and test response capabilities. (EC 2013)

4. The result and discussion

A Nation´s cyber security capability can be understood as a strategic capability in the 21st Century. Building cyber security capability must be based on holistic understanding on different societal areas in order to strengthen overall capability. The measurement of this overall capability is challenging and it is even more challenging to compare cyber security capability between countries. There are two main reasons for that. First, there are several different cybersecurity indices available but they do not measure the same capabilities. At the moment there is not a single cybersecurity index or methodology which we could rely on. Second, specific details of military cyber capabilities, organizations, doctrines and other details often remain hidden from public view, making research particularly difficult. From a nation´s point of view too much detailed information can reveal vital information about what cyber capabilities can do. That can then make it easier for adversaries to defend themselves by blocking the vulnerabilities that these capabilities exploit.

The primary obstacle is that cybersecurity is a sensitive issue, whether from a government or private sector perspective. Admission of vulnerabilities can be seen as a weakness. This is a barrier to the discussion and sharing of threat information and best practices. Yet security through obscurity is not a viable defense model against modern cyber threats. The answer is to implement cybersecurity mechanisms in all layers of society.

Policymakers have a key role to play in ensuring that both public and private entities are well equipped to face the cybersecurity challenges of an ever more connected world. They can achieve this not only by establishing appropriate legal and policy frameworks, but also through promoting cybersecurity awareness and cooperation with the different actors involved in working towards cyber resilience. Developing cyber security in societal level requires strategic understanding of the different elements of cyber security.

The culture of cybersecurity requires collaborative efforts and coordination among all national stakeholders. Effective partnership between public and private sectors is all the more important because non-government entities manage and operate many critical infrastructures that we rely on every day, including those that control transportation, health, banking and energy.

In Finland the Government comprises the highest level of cyber security management. The Prime Minister leads the Government and is responsible for preparing and coordinating the handling of the matters that are the purview of the Government. Preliminary deliberation and coordination occurs in ministerial working groups led by the Prime Minister and, as required, the Government's evening session and negotiations. The Government is responsible for providing political guidance and strategic guidelines for cyber security as well as for taking the required decisions regarding the prerequisites and resources allocated to it.

No single entity or group of stakeholders can secure cyberspace alone — and no individual or group is without responsibility for playing a part in cybersecurity. As not only governments, but organizations of all sizes, as well as consumers, need to take steps to secure their own systems, education and awareness raising play a crucial role. This requires educational and awareness-raising campaigns as well as support for the development and generalization of cybersecurity training in universities and in earlier curricula.

The classifications of the research reports are not entirely consistent with all elements with the DOTMLPF-II. Especially the questions which are related to materiel and facilities are not dealt with in the reports. About the personnel the reports do not deal with quantitative questions, only questions which are related to education and training. Two reports deal with also the legislative questions which the DOTMLPF-II does not consist.

Table 3: Illustrates how the elements of the DOTMLPF-II occur in the four research paper

DOTMLPF-	GSI	EU/BSA	EU/Impact Assessment	Microsoft Security Intelligence
Doctrine	X	X	X	
Organization	X	X	X	
Training	X	X	X	
Materiel				X
Leadership	X	X		
Personnel	X			
Facilities				
Interoperability	X	X	X	

DOTMLPF-	GSI	EU/BSA	EU/Impact Assessment	Microsoft Security Intelligence
Information				X
	Legal	Legal		

The reports do not analyze national cyber security capability comprehensively from a military point of view. The focus is how the nation protects itself against the cyber threats in the peace time and how nation is securing the vital functions of society against cyber threats. Anyway the cyber security capability of the nation gives a very good fundament to build cyber defense capability of the armed forces.

Table 4: Illustrates the Finland rank based on the four cyber security capability research

	GSI	EU/BSA	EU/Impact Assessment	Microsoft Security Intelligence
Finland Rank	8	7	1	1

References

ABI (2015), Global Cybersecurity Index and Cyberwellness Profiles, published 28 May 2015

BSA - The Business Software Alliance (2015), European Union Cybersecurity Maturity Dashboard, A Path to a Secure European Cyberspace, published 9 March 2015

Department of Defense (2015), The Joint Publication (JP) 1-02, Dictionary of Military and Associated Terms, 15 November.

European Commission (2013), Commission Staff Working Document, the Impact Assessment, 7.2.2013, COM 2013 48 final.

Finland's Cyber Security Strategy (2013). Government Resolution, 24 January.

Finland´s Cyber Security Strategy (2013), Background dossier, Security Committee, Ministry of Defence.

International Institute for Strategic Studies (2014). Military Balance 2014. London: Routledge.

Microsoft (2015), Security Intelligence Report, Volume 19, January through June.

The Sound a Rattling Cyber-Sabre Makes: Cases Studies in Cyber Power Projection

Antoine Lemay[1], Scott Knight[2], José Fernandez[1] and Sylvain Leblanc[2]
[1]École Polytechnique de Montréal, Canada
[2]Royal Military College of Canada, Canada
antoine.lemay@polymtl.ca
knight-s@rmc.ca
jose.fernandez@polymtl.ca
sylvain.leblanc@rmc.ca

Abstract: The ability to project power has traditionally been defined as the ability to deploy conventional military assets across the world. While this definition does not apply to a cyber context, cyber forces can still play a role in force projection. By studying the cases of the denial of services attack targeting Estonia in 2007, the Shamoon worm attack of 2012 and the Sony Pictures hack of 2014 as examples of power projection, it is possible to forecast future trends in cyber power projection. Notably, the case studies will show that it is possible to threaten both tangible and intangible assets with cyber attacks. In addition, the case studies will underline the importance of the credibility of the threat for the effectiveness of projection of power, which is influenced both by the perceived capability of the attacker and on the ability to sustain the damage. The case studies also highlight cyber as a means for nation states to project power far outside of their traditional scope, leveraging the loose attributability of cyber force projection for more freedom of action. Based on these observations, the aim of the paper is to anticipate the trajectory of future trends for cyber force projection. First, increased signalling of capabilities would be expected to strengthen the credibility of cyber threats. Second, continued improvement of the ability to apply and sustain cyber attacks is expected to continue due to the increased virtualization of assets of national interest and the increase in pathways for cyber mediated impacts on more traditional assets thanks to the Internet of Things (IoT).

Keywords: cyber warfare, cyber conflict, power projection, cyber strategy, signalling, loose attributability

1. Introduction

As technology develops, the world appears to become smaller. As trains, ships and airplanes have enabled fast travel throughout the globe, Information Technology (IT) has enabled information to reach the other side of the globe in seconds. This technology has brought most nations closer together, but it has also globalized the ability to cause harm. A criminal in Europe or Asia can commit fraud in the United States from the comfort of his living room. With the militarization of the cyber domain, cyber space can be used to conduct military operations almost anywhere in the world. But, can this ability to perform remote operations translate into the ability of a country to impose its will on foreign nations? Our aim is to anticipate the trajectory of future trends for cyber force projection. To begin, we need to evaluate the effectiveness of the cyber domain as a force projection medium.

Because the cyber domain has only recently been identified as a warfighting domain, there is very little publicly available doctrine on how to use cyber power to achieve political or military goals. However, we are gaining access to a number of incidents from which we can learn. We can extract characteristics that will help us better understand cyber as a force projection domain. Studying these cases will also enable us to forecast future trends in cyber power projection in order to start planning for these developments.

The following background section provides information on force projection and how that concept can be interpreted in the cyber realm. Then, three case studies of historic cases of cyber power projection are analyzed in order to extract lessons learned. Finally, the paper uses these lessons to extrapolate possible future trends in cyber force projection.

2. Background

The Dictionary of Military Terms (U.S. DoD, 2009) formally defines power projection as "the capacity of a state to apply all or some of its elements of national power - political, economic, informational, or military - to rapidly and effectively deploy and sustain forces in and from multiple dispersed locations [...]". This definition is inadequate if we ask ourselves whether it would be possible to perform a force projection operation using solely cyber forces. Is the ability to send network packets to remote locations considered sufficient to qualify as

applying informational national power in dispersed locations? To forge a definition of force projection in the cyber context, we must go back to the foundation of power projection.

An important source of conflict between states is diverging national interests. In some cases, it may be possible to find a compromise where both nations can achieve their interests, such as orchestrating a trade or offering compensation. In other cases, such agreement is not possible. In those cases, nations may resort to violence to impose their will. This is what von Clausewitz relates when he states that armed conflict is “a mere continuation of politics by other means” (von Clausewitz, 1832). In this view, the use of force is one of the tools at the disposal of a nation state to pursue its political aims. In other words, a nation can use force (or the threat of force) to compel another nation to pursue a course of action that is against its own interests.

This concept was further studied by Schelling (Schelling, 1960) who used game theory in the context of nuclear deterrence to model this effect. The game theoretical approach models each nation's incentives to pursue courses of actions based on the possible courses of action pursed by the other nation, providing a mathematical model of the nation's political calculations. The hypothesis is that a rational decision maker will select the course of action that is the most advantageous. Using this approach, we can create a general concept of force projection. By applying negative incentives (e.g. by the use of force) to the choice of a course of action, it is possible to alter the behaviour of rational actors and incite them to choose a different course of action. For example, if nation A stands to gain 1 billion dollars by protecting its market and nation B sends its navy and creates 10 billion dollars worth of damage, the net effect for nation A to protect its market is a 9 billion dollars loss, which is a sub-optimal decision for nation A. In this example, the ability to apply this hard power has altered the calculations of nation A to the benefit of nation B.

Ideally, one would want the benefit without having to actually apply hard power. Announcing that hard power will be applied if a given course of action is selected would cause a rational target to factor the resulting costs into their calculations and adjust their choice accordingly. Multiple examples of this can be found in the history of nuclear strategy (Freedman, 2003), such as the role of nuclear deterrence during the Cold War. Because of the threat of force, i.e. the menace of a nuclear war, both the U.S. and the U.S.S.R. altered their political calculations, be it by removing missiles from Cuba or by preventing General McArthur from escalating the Korean conflict. As such, the ability to threaten the seizure or the destruction of foreign assets through “the capacity of a state to apply all or some of its elements of national power [...] in and from multiple dispersed locations” can be considered an important function of hard power to achieve political objectives. The ability to threaten in such a manner through the cyber domain is what we will call cyber power projection (or cyber force projection) in the context of this paper.

The concept of using the cyber realm to perform military actions is not new. Even the U.S. the Department of Defense (DoD) Law of War Manual (Office of the General Counsel DoD, 2015) contains a chapter on cyber operations. However, there does not seem to be a public domain study of how these tactical cyber operations link back to the strategic national interests. The DoD Cyber Strategy document (DoD, 2015) does shed some light of the link between cyber assets and strategic interests. However, it is mainly focused on deterring cyber attacks from foreign states. It shows that the DoD has deemed the cyber domain a viable force projection medium for other states, but this sheds little light of what form this cyber power projection would take. Thus, we must look elsewhere to gain insight.

3. Case studies

This section will study three cases of large scale cyber attacks in the context of their effectiveness as a tool for imposing national will abroad. First, the case of the cyber attacks against Estonia is presented. This is followed by the case of the Shamoon worm. Finally, the campaign against Sony Pictures is discussed.

3.1 Estonia 2007

In April 2007, following the relocation of a Soviet war memorial, Estonia was hit with a massive coordinated denial of service attack lasting several days (Kostadinov, 2013). This attack knocked many government websites offline, but also affected other services, such as bank ATMs. In addition to the denial of service attacks, multiple web sites were defaced to oppose the move and to present a negative view of the Estonian government. Service was eventually restored by Estonia's cyber defenders and the attacks abated.

While there has been no publicly disclosed evidence linking the incident to the Russian Federation's state apparatus, it is clear that the Russian state used the incident to put pressure on Estonia. President Putin commented on the affair during the May day parade, lamenting the desecration of memorials to war heroes (Davis, 2007), and people close to the government, including an activist from the government-backed youth organization Nashi (Shachtman, 2009) and an aide to a member of the Russian Duma (Coalson, 2009), have claimed responsibility for organizing the attacks. In this sense, we can argue that the Estonian cyber attacks represent an example of leveraging of the Russian state's ability to project cyber attacks outside their national boundaries for political gain, even if the attack was not directly launched or coordinated by their armed forces. This argument is further strengthened by the application of more traditional forms of pressure, on the Estonian government, such as the suspension of train services linking Russia and Estonia.

Thus, we can argue that the cyber attacks appear to be what von Clausewitz would dub an "extension of politics by other means". In the case of Estonia, the political calculations appear clear; as long as the Estonian government pursued its course of action of relocating the war memorial, it would continue to feel the pain. Instead of caving in to demands, however, the Estonian government chose to ease the pain by fighting off the cyber attacks. It invoked NATO's mutual assistance rule and put in place technical measures to repel further attacks and limit damages. This hampered the ability to sustain political pressure through the cyber attacks and thus removed its viability to coerce Estonia and may even have further distanced Estonia from the Russian sphere of influence.

By analyzing their outcome, we could argue that the 2007 cyber attacks against Estonia were not an effective example of force projection. It was not possible to alter Estonia's course of action to one preferred by Russia. This lack of success appears to stem from the ability of Estonia to mount a credible defence and mitigate the impact of the cyber attack. The natural implication for future use of cyber power as force projection is that the ability to sustain the attack is critical. In other words, the ability to deliver on the threat of pain is crucial to the credibility of the threat. If the threat is not credible, then the cost of damages can be removed from the political calculations, thus rendering the cyber power ineffective as a way to alter foreign policy. It is interesting to note that the fact that the origin of the attacks could not be attributed with certainty to the Russian government did not diminish the credibility of the implied threat. In fact, this *loose attributabilty* may have increased the credibility of the threat because without solid attribution the normal diplomatic cost associated with aggression may be avoided. This can make the use of such threats more attractive and more likely to be employed.

3.2 Shamoon 2012

In August 2012, the computer network of Saudi Aramco, the oil company of Saudi Arabia, was infected by the *Shamoon* worm. The worm spread throughout the network, affecting close to 30,000 machines, and even spreading to RasGas (the Qatari gas company) and ExxonMobil (Bronk and Tikk-Ringas, 2013). The Shamoon contained a Wiper component that erased the contents of the hard drive, rendering the affected machines unbootable, and leaving them displaying an image of a burning American flag (Symantec Security Response, 2012). While the damage could have been worse if the control systems for oil production had been affected, production was still disrupted for a few days due to efforts deployed to respond to the incident.

As with the Estonian attacks, exact attribution of the Shamoon worm is difficult. A group called "The Cutting Sword of Justice" has claimed authorship of the Shamoon attack, supposedly in retaliation for "crimes and atrocities" in neighbouring countries. In fact, U.S. officials have identified Iran as the sponsor of the attack (Mount, 2012) and NSA documents recently exposed by Edward Snowden, implying that Iran learned from cyber attacks against its own infrastructure (e.g. Stuxnet) to create the Shamoon worm, seem to strengthen this hypothesis (Zetter, 2015). Based on this information, it seems likely that the state of Iran supported or at the very least encouraged these attacks.

Under the assumption that the Shamoon attack is indeed an example of Iranian cyber force projection, we can wonder what political objectives were being pursued. Bronk and Tikk-Ringas argue that the principal motive was to damage and exact revenge on Saudi Arabia, a historic enemy, by jeopardizing their principal revenue stream, oil production. They also point out that this option was particularly enticing at the time, when economic sanctions severely limited their oil exports, further incentivizing Iran to drive the price of oil up. This line of thought is particularly interesting if we consider that oil production capability is a tangible asset, not a virtual one. A cyber attack could affect oil production more cost efficiently than physical means. For example, at the

same time, Iran was threatening to close the straits of Hormuz, an ineffective threat given the presence of significant U.S. naval forces in that theater.

Other political objectives may however have been pursued in the Shamoon attack. The attack may have been used to deter the United States from further cyber attacks on Iran. It has been widely assumed that the US and Israel were responsible for the Stuxnet and Flame cyber attacks on Iran. By demonstrating their ability to successfully deliver a cyber attack abroad, i.e. by demonstrating their ability to project force in cyber space, Iran may have wanted to signal that it could retaliate in kind. The success of the Shamoon operation lends credibility to the cyber retaliation threat, credibility that Iran does not possess in the physical domain. The threat of a Shamoon-like event occurring in the continental United States is much more credible than an Iranian physical attack, and acts as a better deterrent. Also, a critical component of the Shamoon malware has the same name, “wiper”, and the same behaviour as a component of the Flame malware signalling that further attacks would strengthen Iranian capabilities. The credibility of the threat is highlighted by the concern expressed on precisely this point in the leaked NSA documents (Zetter, 2015)

By looking at the Shamoon incident, we can see that Iran significantly increased its geographical reach through cyber space. Through cyber space, Iran can damage physical assets (Saudi oil production capacity) that are out of reach of its conventional military and unconventional forces such as Hezbollah. Furthermore, it might allow Iran to project enough force to deter U.S. actions, which could not be achieved through military force by a regional power like Iran. This seems to imply that developing the ability to project cyber power can be used to project force globally without the need for traditional force projection apparatus, such as a blue-water navy or ballistic missiles. It is also interesting to note that this is another example of loose attributability. In this instance Iran needed to conduct a successful, destructive attack to establish credibility, but it did so through a fronting group, the "Cutting Sword of Justice". By unattributably signalling that it is responsible for Shamoon, Iran may be establishing the credibility of its cyber capability and its willingness to use it in deterrence.

3.3 Sony Pictures 2014

As described in an security industry report in November 2014 (RiskBased Security, 2014), a group called Guardians of Peace (GOP) hacked into the Sony Pictures network and exfiltrated a trove of confidential information, including unreleased movies, employee personal information, internal emails and even details of secret business deals. The group then started to release information piece by piece, offering to stop if the release of the movie “The Interview”, a story about assassinating North Korean dictator Kim Jung Un, was cancelled. Sony Pictures refused to comply with the demands even as more damaging information was released to the public. Ultimately, the GOP group released a statement threatening terror attacks on theatres showing the movie. This prompted several large theatre chains to remove the movie from their program. The potential loss of revenue convinced Sony Pictures to cancel the theatrical release, instead opting to release it in an on-demand format.

Unlike the previous case studies, this attack was very confidently attributed to North Korean actors by the United States and sanctions were levied (Lee and Solomon, 2015). While this claim is very hard to validate independently given that the sources for this attribution are not public, the lack of protest to US retaliation by the international community seems to indicate consensus on this question. North Korea’s aim appears to have been to coerce Sony Pictures to cancel the release of a movie picturing North Korea's leader in an unfavourable light. Much like the Estonian case, cyber force was projected to cause harm for as long at the target did not comply with a desired course of action. In this case, however, even if the threat showed high credibility because of the data already exfiltrated, Sony Pictures did not give in to demands. As such, cyber force projection backed by a credible threat proved ineffective in achieving the political objectives of North Korea.

When the application of cyber force proved unsuccessful, physical terror attacks were threatened. We could argue that these threats were largely lacking in credibility as North Korea possesses limited ability to project conventional power outside its immediate sphere of influence and no history of use of non-conventional forces. However, these threats proved to be very effective, convincing multiple theatre chains to pull the movie. We can speculate that the successful delivery on the cyber threat greatly enhanced the credibility of this second threat. This forces us to consider not only the credibility of the threat itself, but its perceived credibility in the eyes of the target, no matter how factual it actually is.

It is also important to note the severity of the damage inflicted on Sony Pictures, which spent more than 15 million dollars in investigation and remediation costs (Musil, 2015), forced the resignation of executives, lost sales from leaked movies, saw damages to its brand, etc. Of note is the fact that this financial damages are accrued by the loss intangible assets such as intellectual property and other confidential information. This serves as a reminder that, as technology progresses, more of the valuable assets that we care about reside in cyber space. As such, there is no reason to believe that a threat to these assets would be less effective as a way to coerce foreign actors than a threat to a more conventional asset of the same value. Based on this observation, it is likely that the ability to project force in cyber space will increase and become a staple of power projection, especially in cases where it is difficult to affect intangible assets through conventional force.

4. Expected future trends

Looking back at these three case studies, it is clear that the credibility of the threat is a critical factor in the use of cyber power projection as a means to achieve political objectives. A foreign power will not be coerced into taking a course of action that is against its interest if it does not believe that harm could come to it. In that sense, we can expect that a reinforcement of the credibility of the threat will be important.

Stuxnet is another example of cyber force projection to achieve an effect. Previously, the Israeli Air Force had used physical attacks to deter its neighbours from developing nuclear weapons. Stuxnet achieved similar goals by disabling large portions of an Iranian uranium centrifuge plant, without direct diplomatic fallout due to the difficulties in attribution of cyber attacks. Following the discovery of Stuxnet and its public investigation, there was a significant leak of details surrounding an American-Israeli program called Olympic Games, the Stuxnet operation (Sanger, 2012). Although never officially substantiated, the leak was never aggressively denied (loose attributablility again). Whether the leak was intentional or not, it signals deep technical capability and a willingness to employ the capability to achieve policy outcomes. This establishes credibility for American cyber force projection.

A more recent example of a cyber attack that may have had similar effects is the December 2015 attack on the Ukrainian electrical grid. The coordinated attack caused significant power outages (Lee, 2016). A well-known malware called BlackEnergy was used in the attack, and although there is no confirmed attribution it is thought to have been deployed by a Russian hacking group called Sandworm, which has been tied to Russian government interests (Rashid, 2014). Whether this attribution is valid or not, the result is that it strongly signals the credibility the Russian cyber threat. The attack was very well coordinated and effective, indicating deep capability. At a time when it is more difficult for Russia to provide direct military support to Ukrainian separatists due to geopolitical constraints, the attack keeps pressure on the Ukrainian government and signals support to the separatist cause. It further signals a willingness to use cyber attacks in force projection. It is accomplishing these influence goals with little diplomatic fallout because the source of the attacks cannot be proven. Nonetheless, the credibility of cyber force projection is enhanced because of loose attributability.

Based on our observations, we believe that increased signalling of capabilities is to be expected in the future. However, the main problem with advertising cyber capability is that preserving the secrecy of the method may be required to maintain the capability. For example, an attacker that can project force anywhere because it possesses a 0-day vulnerability loses its capability once it is used or disclosed and a patch can be created. For example, consider the Shamoon worm. Iran managed to send a clear signal to the United States that it had the know-how and operational capacity to launch attacks against critical infrastructure, without needing to divulge any specifics of its cyber program, especially given that may have reused technical components of the American cyber program.

The flip-side of the increased virtualization of assets is the increased ability to reach physical assets through cyber space, as illustrated by the Shamoon case study. The fact that it was possible for Iran to affect oil production through cyber space demonstrates the ability that some actors have to use cyber space as a medium for more conventional force projection through cyber-mediated attacks. While certain countries speak of and strive to achieve *cyber dominance*, as of yet no country can pretend to have the achieved the ability to reliably interdict cyber-mediated attacks. As such, it is likely that the cyber domain will remain a way to build a global force projection capability on a limited budget, even if only targeting virtualized assets or those physical assets that can be reached through cyber space.

In most industrialized countries, the assets that can be targeted through cyber mediated attacks are already significant. Most critical infrastructure and manufacturing heavily leverage industrial automation, for example SCADA networks, to operate. This exposes that infrastructure to cyber-mediated attacks, thus making them attractive targets. Furthermore, this trend will continue to accelerate with the development and deployment of new technologies such as automated cars, delivery drones, networked insulin pumps and pacemakers. All these physical devices, which contain embedded computing, connecting to the global communication networks, what has been dubbed the *Internet of Things* (IoT), provide new targets for cyber-mediated attacks and further increase the value of cyber force projection. In that light, we can expect more actors to invest in this capability.

5. Conclusion

The ability to project power is essential to leverage military power to coerce or deter foreign nations. Because of the global reach of cyber space, the cyber domain seems like a natural medium to project power. By analyzing the cases of the Estonian cyber attacks, the Shamoon worm and the Sony Pictures hack, we have substantiated that observation, but with a number of caveats. First, we have seen from the case of Estonia that if the defender does not believe that the aggressor can sustain the threat, cyber power projection has limited coercion capability. This credibility problem has also surfaced in the Sony Pictures case where the first threat was not taken seriously, but subsequent threats, even if they were objectively less believable than the first, achieved the desired effect because of the credence obtained in the highly publicized attack.

Because the credibility of the cyber threat appears to be so important, we have speculated that the current policy of secrecy will not be sustainable and that we will observe more signalling of cyber capabilities in the future. Credible cyber force projection relies both on an adversary's technical capability and on his willingness to use that capability. When employed for that purpose, the loose attributability of cyber attacks provides a double advantage to cyber force projection with respect to other traditional forms of force projection. On the one hand, public displays of support, circumstantial technical evidence and orchestrated leaks can be employed to let the victim of the attacks suspect who the originator is, thus achieving aims of credibility and deterrence. On the other hand, the fact that retaliation to cyber attacks without substantial proof of attribution could often be damaging provides its perpetrator a certain level of impunity.

Another significant advantage of cyber power projection is that it is possible to reach both physical and intangible assets through cyber operations that would be otherwise unreachable by more traditional means. In the Shamoon case study, oil production capacity, a physical asset, was affected. The case of Sony Pictures demonstrated the increasingly strategic significance of intangible assets. The study of the Shamoon worm has also demonstrated that a nation with limited ability to project power in the physical realm might effectively project power in cyber space.

The combination of these advantages greatly enhances the reach of nations that do not possess more traditional power projection capabilities or could not afford the negative geopolitical consequences of physical power projections. While the cases we studied do not cover cyber attacks exhaustively, our observations allow us to predict a trend toward more cyber attacks used as tools for force projection. Such future attacks will surely shed more light on the use of cyber power projection. One remaining open question is whether future users of cyber force projection might be able to effectively signal capabilities while preserving the secrecy of their methods. Better signalling could significantly reduce uncertainty in calculations from the target of force projection and limit the risks of accidental escalation.

References

Bronk, C. and Tikk-Ringas, E. (2013) "Hack or attack? Shamoon and the Evolution of Cyber Conflict" in *Survival, Global Politics and Strategy*. [Available Online] http://bakerinstitute.org/media/files/Research/dd3345ce/ITP-pub-WorkingPaper-ShamoonCyberConflict-020113.pdf [Accessed 4 February 2016]

Coalson, R. (2009) "Behind The Estonia Cyberattacks". Radio Free Europe Radio Liberty. [Online] http://www.rferl.org/content/Behind_The_Estonia_Cyberattacks/1505613.html [Accessed 4 February 2016]

Davis, J. (2007) "Hackers Take Down the Most Wired Country in Europe". Wired Magazine, Issue 15.09. [Online] https://archive.wired.com/politics/security/magazine/15-09/ff_estonia?currentPage=all [Accessed 4 February 2016]

Department of Defense (2015) *The DoD Cyber Strategy*. April 2015. [Available Online] http://www.defense.gov/Portals/1/features/2015/0415_cyber-strategy/Final_2015_DoD_CYBER_STRATEGY_for_web.pdf [Accessed 4 February 2016]

Kostadinov, D. (2013) "Estonia: To Black Out an Entire Country – part one", Infosec Institute. [Online] http://resources.infosecinstitute.com/estonia-to-black-out-an-entire-country-part-one/ [Accessed 4 February 2016]

Lee, R. M. (2016) "Confirmation of a Coordinated Attack on the Ukrainian Power Grid". SANS Industrial Control Systems Security Blog. [Online] http://ics.sans.org/blog/2016/01/09/confirmation-of-a-coordinated-attack-on-the-ukrainian-power-grid?utm_source=hs_email&utm_medium=email&utm_content=25135530&_hsenc=p2ANqtz-87XLhYBXFcESdxOIJIB8DSoYBZ5sPrfHQv9xNUp11BwFsfcUBouRDj-R7y6YcJY2BsrUeKvRVbwO4lPcVAPgHLmDrj7w&_hsmi=25135530 [Accessed 4 February 2016]

Lee, C. E. and Solomon, J. (2015) "U.S. Targets North Korea in Retaliation for Sony Hack". The Wall Street Journal. [Online] http://www.wsj.com/articles/u-s-penalizes-north-korea-in-retaliation-for-sony-hack-1420225942 [Accessed 4 February 2016]

Mount, M. (2012) "U.S. Officials believe Iran behind recent cyber attacks", CNN. [Online] http://www.cnn.com/2012/10/15/world/iran-cyber/ [Accessed 4 February 2016]

Musil, S. (2015) "Sony Pictures hack has cost the company only $15 million so far". CNET. [Online] http://www.cnet.com/news/sony-pictures-hack-to-cost-the-company-only-15-million/ [Accessed 4 February 2016]

Office of the General Counsel Department of Defense (2015) "Cyber Operations", in *Department of Defense Law of War Manual*. pp. 994-1009.

Rashid, F. Y. "Russia-linked Hackers Exploited Windows Zero-day to Spy on NATO, EU, Others". Security Week. [Online] http://www.securityweek.com/russian-hackers-exploited-windows-zero-day-spy-nato-eu-other-high-profile-targets [Accessed 4 February 2016]

RiskBased Security, (2014)"A Breakdown and Analysis of the December, 2014 Sony Hack". RiskBased Security. [Online] https://www.riskbasedsecurity.com/2014/12/a-breakdown-and-analysis-of-the-december-2014-sony-hack/ [Accessed 4 February 2016]

Sanger, D. E. (2012) "Obama Order Sped Up Wave of Cyberattacks Against Iran", New York Times. [Online] http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html [Accessed 4 February 2016]

Schelling, T. C. (1960) *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Shachtman, N. (2009) "Kremlin Kids: We Launched the Estonian Cyber War". Wired. [Online] http://www.wired.com/2009/03/pro-kremlin-gro/ [Accessed 4 February 2016]

Symantec Security Response (2012) "The Shamoon Attacks". Symantec Official Blog. [Online] http://www.symantec.com/connect/blogs/shamoon-attacks [Accessed 4 February 2016]

The United States Department of Defense (2009) *The Dictionary of Military Terms*. New York, NY: Skyhorse Publishing.

von Clausewitz, C. (1832) *On War*.

Freedman, L. (2003) *The Evolution of Nuclear Strategy - Third Edition*. London, U.K. : Palgrave Macmillan.

Zetter, K. (2015) "The NSA Acknowledges What We All Feared: Iran Learns From US Cyberattacks", Wired. [Online] http://www.wired.com/2015/02/nsa-acknowledges-feared-iran-learns-us-cyberattacks/ [Accessed 4 February 2016]

Exploring the Puzzle of Cyberspace Governance

Andrew Liaropoulos
University of Piraeus, School of Economics, Business and International Studies, Department of International and European Studies, Piraeus, Greece
aliarop@unipi.gr
andrewliaropoulos@gmail.com

Abstract: Cyberspace is a socio-political and technological domain with unique characteristics. It is a human made domain that offers universal reach and access to its users. Its decentralized nature and the fact that it is mostly owned and managed by the private sector, raise a number of questions regarding the limits of state sovereignty and the most effective form of governance. Viewing cyberspace as a global commons, balancing between state sovereignty and the fragmentation of cyberspace and debating between multilateral governance and multi-stakeholderism, illustrate the difficulty of regulating human activities behind keyboards and computer screens. The cases of ITU, ICANN, IGF and NETmundial offer us a pragmatic insight into the power politics of cyberspace. Reducing uncertainty between the various stakeholders, developing norms, advancing law-making efforts and matching geopolitics with technology are all pieces of the complex puzzle of cyberspace governance.

Keywords: cyberspace governance, state sovereignty, multilateral governance, multi-stakeholderism, minilateralism

1. Introduction

Cyberspace is described as the 'environment formed by physical and nonphysical components characterized by the use of computers and the electromagnetic spectrum to store, modify and exchange data using computer networks' (Boothby 2014, 123). Cyberspace is a global digital network that is embedded in every aspect of our daily life. It encompasses not only Internet, but also the critical infrastructure that supports modern societies, like the electrical grids, water supply systems, banking transactions and transportation systems. Over the past two decades, the rapid evolution of cyberspace has affected the way societies communicate and interact in the political, economic and social sphere. Cyberspace is a channel of globalization; it is universal and open to all. The communication and information exchange that it provides has virtually reduced the size of the world. But as with any other channel of communication, cyberspace is not immune to insecurity, crime and competition. Cases of cyber-espionage, data losses, compromised networks and cyber-doom scenarios fill the headlines on a daily basis. Cyberspace is an integral part of the critical infrastructure on which governments, the private sector and citizens depend. States, international organizations, private companies and human rights activists are struggling to regulate a wide range of activities that take place in cyberspace and at the same time balance between critical infrastructure protection, civil liberties, technical standards and cost.

The governance of cyberspace is a complex task. Due to its asymmetrical, anonymous and dual-use features, cyberspace challenges traditional understanding of key concepts like security, borders, human rights, privacy and sovereignty (Slack 2016). The reason is that the socio-political and technological characteristics of this new domain are constantly being redefined (Choucri 2012, 4).[1] The rapid pace of technological change and the way societies respond in the digital realm is affecting the interests of state and non-state actors in cyberspace. Advances in the field of information technology, like the Internet of Things (Weber 2013),[2] Big Data (Cukier & Mayer-Schoenberger 2013)[3] and the Dark Web (Chertoff & Simon 2015),[4] have surpassed the ability of states and international organizations to offer efficient governance. Cyberspace is largely owned and managed by the

[1] According to Professor Nazli Choucri, cyberspace is characterized by: temporality (replaces conventional temporality with near instantaneity), physicality (transcends constraints of geography and physical location), permeation (penetrates boundaries and jurisdictions), fluidity (manifests sustained shifts and reconfigurations), participation (reduces barriers to activism and political expression), attribution (obscures identities of actors and links to action) and accountability (bypasses mechanisms of responsibility).
[2] The Internet of Things (IoT) is a concept that aims to connect various devices or objects - things through wireless and wired connections and create an environment where users can interact at any time with the digital and the physical world. The IoT is growing rapidly. Mobile applications and sensors are now operating in cars, refrigerators, machinery, medical technology and smart phones.
[3] Big Data is a term that refers to large and complex sets of data, that surpass the ability of typical database software tools to capture, store, manage and analyze. The challenges that Big Data poses, relate to the '3Vs' characteristics: volume, variety and velocity.
[4] Dark Web is a part of Internet that is intentionally hidden; it is not indexed by search engines and is inaccessible through standard web browsers. An example of Dark Web is the Tor network that offers its users anonymity by encrypting data and sending them through other routers.

private sector, but is increasingly affecting civil society and challenging state sovereignty. This reality poses a great challenge to the traditional idea of global governance that is mainly state-centric.

The purpose of this article is to shed light on the complexity of cyberspace governance. The first section briefly analyses the idea of global governance and addresses the implications that cyberspace poses to state sovereignty. The challenges of applying the Westphalian concept of sovereignty in the seemingly borderless cyber domain, are reflected in the analysis of the multilateral governance model and multi-stakeholderism. The second section critically reviews these two approaches in the cases of the International Telecommunication Union (ITU), the Internet Corporation for Assigned Names and Numbers (ICANN), the Internet Governance Forum (IGF) and the Global Multistakeholder Meeting on the Future of Internet Governance (NETmundial), in order to identify the advantages and disadvantages of each approach. The final section aims to bring together the pieces of the puzzle, and address the need for an alternative approach towards cyberspace governance, that of minilateralism.

2. Cyberspace, state sovereignty and global governance

Definitions about cyberspace abound. It is common to view cyberspace in three layers. The first one is the physical layer that consists of electrical energy, integrated circuits, communications infrastructure, fiber optics, transmitters and receivers. The second layer is the software, meaning the computer programs that process information. The last and least concrete layer is that of data (Tabansky 2011, 77-8). It is important to distinguish between the physical and non physical aspects of cyberspace. Gourley differentiates between the *domain* (the medium) and the *space*. He refers to the physical aspects as the *cyber domain* - anything that enables users to transmit, store and modify digital data. Cyber activities take place *through* the cyber domain *in* cyber space[5] (Gourley 2014, 278). The cyber domain is an artificial and human-made construct with geographical ties over a specific territory. Thereby, states can exercise their sovereignty through the cyber domain. According to Gourley the territoriality principle allows states to control cyber activities occurring within and across their borders and the effects principle gives them jurisdiction over external activities that cause effects internally (279). Applying the same analogy for cyber space (the non physical aspects of cyberspace) is tricky. There is no common approach regarding sovereignty *in* cyber space. States differ on whether cyber space is a global commons, on how sovereignty can be exercised in cyber space due to the attribution problem and on whether it is in their national interest to exercise sovereignty at the first place. The discussion regarding sovereignty in cyber space and the efforts to recognize sovereignty over cyber space are critical in understanding the dilemmas of governance (Gourley 2014 279-280).

The concept of governance refers to the governmental institutions and informal regulatory mechanisms that guide and restrain the collective activities of a society. Governance illustrates a system of governing methods where the boundaries of public and private sectors are unclear. Governance has a wider meaning than government. The latter is an executive apparatus that can exist in the presence of widespread opposition to its policies, whereas the former requires acceptance by the majority of those it affects. The term governance is a rather fuzzy term that has been used in a variety of ways in the international relations literature. Global governance does not refer to the creation of a global government, but to the cooperative efforts of states, international organizations and non-state actors to address common challenges that transcend national borders (Patrick 2014, 59).

In brief, the main arguments in the literature on global governance are the following (Nye & Donahue 2010; Rosenau 1995; Rosenau & Czempiel 1992). First, that there is a shift to regulation from the national level, to levels beyond the state. Second, that world politics is more than just intergovernmental politics and that the areas of authority beyond the state have increased. Finally, that rules beyond the state are legitimate, if the representatives of affected interests have agreed upon them in a decision-making process that meets reasonable standards of inclusiveness, transparency and accountability (Dingwerth 2008). Global governance is not conducted exclusively by governments and international organizations, but also by the private sector and non-governmental organizations (NGOs). As a result, states are not replaced as the primary instrument of global governance, but rather supplemented by other actors (Nye & Donahue 2010, 12).

When approaching cyberspace governance, we should consider several issues (Cornish 2015; Deibert 2013; DeNardis 2014; Jayawardane, Larik & Jackson 2015; Nye 2014; West 2014). Should cyberspace be governed?

[5] Note that Gourley uses the term *cyber space* (two words) in order to differentiate from cyberspace.

Who should be involved in governance? How should cyberspace be governed? How far should regulation go? There are three main approaches that address the above issues: distributed governance, multilateral governance and multi-stakeholderism (West 2014, 4). In the early days of Internet development, governance could be described as a distributed system. Governance was limited, unorganized and restricted within online communities, who asserted that information had to be free, and not controlled (Deibert & Crete-Nishihata 2012, 341-342). This approach reflected an era where online communities where small, homogeneous and able to regulate themselves. In 1996, John Perry Barlow, the founder of the Electronic Frontier Foundation (EFF)[6] stated in *The Declaration of the Independence of Cyberspace* that 'Governments are not welcome among us…Cyberspace does not lie within your borders…We are forming our own Social Contract. This governance will arise according to the conditions of our world, not yours' (Barlow 1996). In the 1990s, Internet had less than a million users and was in a primitive phase of development. Nowadays, the Internet users are counted in billions and cyberspace has become an integral part of modern societies (Betz & Stevens 2011, 15). Cyberspace has matured and it has reached an evolutionary phase where regulations are needed. The distributed governance model, although still popular in some online communities, cannot provide efficient policy solutions that are acceptable to the large and diverse community of cyberspace users.

The argument that state sovereignty should have a limited role in cyberspace has also been embraced, by those who view cyberspace as a global commons. In sharp contrast to land, sea, air and space, cyberspace is a human-made domain that lacks physical space and thereby borders. Cyberspace comprises a global common infrastructure, but is not a global commons (Cornish 2015, 158). Cyberspace seems borderless, but is actually bounded by the physical infrastructures that facilitate the transfer of data and information. Such infrastructures are mostly owned by the private sector and are located in the sovereign territory of states. There is no doubt that states are trying to overcome the so-called border paradox and develop virtual borders (Demchak & Dombrowski 2011). James Lewis eloquently described cyberspace as a condominium, with many owners (2010, 16). Paul Cornish labels cyberspace as a *virtual commons* that is neither private property, nor sovereign territory, nor global commons in the same way that sea and the air are considered to be (2015, 158-159).

The issue of state sovereignty is of central importance to the advocates of the multilateral governance. The multilateral approach views cyberspace in Hobbesian terms. Supporters of this state-centric approach understand cyberspace as a chaotic domain that reinforces insecurity and therefore argue that states should be the ones to formulate policy in cyberspace. This approach calls for the creation of a body within the United Nations (UN), that will be responsible for cyberspace governance, but at the same time states will have the power to set their own national policies. The multilateral model has traditionally been supported by Russia, China, India, Iran and Saudi Arabia. In the aftermath of the Edward Snowden disclosure, it has gained momentum even among some EU member states that seek to protect their cyber-borders and data from the surveillance systems of the US (West 2014, 7). National governments view the privacy policies adopted by transnational companies like Google, Facebook and Twitter as a threat to digital sovereignty and thereby national security (Nocetti 2015, 114). It has been argued that in an era of great power antagonism, the exercise of state sovereignty in order to secure national digital assets and critical infrastructure, could lead to the fragmentation - *Balkanization* of cyberspace. Cornish points out whether fragmentation is actually a negative development and should be considered a threat to cyberspace or rather a credible alternative to it (2015, 159). States have the choice to break off from the current Internet and form their own regional or national Intranets. States are examining the option of creating 'national cyberspaces', building trans-oceanic cables and store Internet data on servers within their national territories, but still prefer the economic benefits of connectivity.

For the advocates of the multilateral approach, Internet governance should respect the Westphalian notion of sovereignty and therefore should resemble the case of the International Telecommunication Union.[7] The protection of digital sovereignty and information security are the main priorities for states that embrace the multilateral governance model. Another example that fits to a certain extent with the multilateral model is the Shanghai Cooperation Organization (SCO). Russia, China, India, Iran and other Central Asian states have been coordinating their Internet security policies through the SCO and conducting cyber-exercises designed to counter Internet-enabled political uprisings. SCO is an example of low-level multilateral cooperation between states that prefer a tightly controlled Internet (Deibert 2015, 13).

[6] For more details see https://www.eff.org/.
[7] For more details see http://www.itu.int/en/Pages/default.aspx.

In sharp contrast to the above, the multi-stakeholder governance model involves state and non-state actors that represent the business sector and civil society. The rationale is that governments alone cannot regulate cyberspace successfully. Therefore, other actors like technical corporations, search engines, internet users and civil organizations should also be involved in the governance of Internet. Microsoft, Apple, Google, Yahoo, Weibo, Skype, Dropbox, Amazon, Twitter, Facebook and Badoo are only some of the numerous companies, technical providers and search engines that collect and store data. The advocates of the multi-stakeholder governance model argue that cyberspace norms will be accepted by internet users, only if they are part of designing them. This will enhance legitimacy and authority of institutions, organizations and companies in cyberspace (Mihr 2014). Supported by the US, UK, Canada, Australia and organizations like Google and ICANN, the multi-stakeholder model has been quite popular in the pre-Snowden era. In the aftermath of the Snowden disclosure the legitimacy and credibility of this approach has been considerably weakened (Deibert 2015, 13). The next section will apply the multilateral governance and multi-stakeholderism in the cases of ITU, ICANN, IGF and NETmundial.

3. Cyberspace governance: multilateral governance and multi-stakeholderism

The case of the ITU is a perfect example not only of the challenges that Internet governance is facing, but also of the *battle* between multilateral governance and multi-stakeholderism. The ITU is a UN body responsible for international telecommunications and is regarded, especially by the least developed countries, as the most appropriate forum for the governance of Internet (Jayawardane, Larik & Jackson 2015, 6). Even though the ITU lists 700 private sector entities among its membership, it is not considered a good example of multi-stakeholderism. The reason is that only the 193 member states that participate and vote in the Plenipotentiary Conference, can decide on the future policies of the organization. The ITU is therefore, a multilateral organization, where only states can formulate policy (Glen 2014, 637). The limited participation of civil society organizations in the ITU and the attempt during the World Conference on International Telecommunications (WCIT) in December 2012 to transfer responsibility for Internet governance from bodies such as ICANN to the ITU, represent a major challenge to multi-stakeholderism (Glen 2014, 651). Over the past years, states attempt to territorialize cyberspace and replace multi-stakeholderism with a centralized and multilateral model (Glen 2014). This trend gained further momentum after the revelations by Edward Snowden regarding the cyber surveillance programs conducted by the National Security Agency (NSA).

Although multi-stakeholderism is considered nowadays the mainstream approach in Internet governance, it was only on 2002 that the United Nations General Assembly (UNGA) identified the role of other participants, apart from states, in safeguarding cyberspace. In particular, the UNGA Resolution 57/239 of 2002 made reference to 'governments, businesses, other organizations and individual users who develop, own, provide, manage, service and use information systems and networks'. According to the resolution, the participants 'must assume responsibility for and take steps to enhance the security of these information technologies, in a manner appropriate to their roles' (Kremer & Müller 2014, 15). The term 'stakeholders' first appeared in the UNGA Resolution 58/199 of 2003. In 2010 the Report A/65/201 of the UN Governmental Group of Experts (GGE), stressed the importance of 'Collaboration among states, and between states, the private sector and civil society', thereby recognizing an equal role for civil society in the governance of cyberspace (Kremer & Müller 2014, 15).

Multi-stakeholderism advocates the inclusive participation of all relevant actors that deal with cyberspace governance. These actors include not only states, but also a variety of non-state actors, like civil society groups, representatives of the private sector, media and other actors that regulate communication in cyberspace. The advantage of the multi-stakeholder approach is that all relevant actors can participate and be heard on an equal basis (Mihr 2014, 28). Inclusiveness and representativeness are the core principles of this approach. In an ideal scenario, the stakeholders do not only produce norms and set their own standards, but also define possible repercussions or penalties for non-compliance (Mihr 2014, 28). Multi-stakeholderism should not be understood as an end in itself, but rather as a process to reach effective governance. Multi-stakeholderism cannot and does not aim to replace states. Besides, stakeholders do not all participate in the same way and to the same extent in the governance of cyberspace. For example, civil society actors, private sector organizations and global think tanks might play a leading in shaping and institutionalizing norms of behavior in cyberspace, but it is only states that can enforce regulations (Jayawardane, Larik & Jackson 2015, 4-5). In order to further explore this approach we will briefly analyze the most influential cyberspace governance fora. These include ICANN, IGF and NETmundial. Each case will provide us with different insight on the uses and limits of multi-stakeholderism.

When the Internet expanded globally in 1997, the US government created ICANN.[8] The later is an internationally organized nonprofit organization that is responsible for the Internet Assigned Numbers Authority (IANA) functions, mainly Internet Protocol (IP) space allocations, the Domain Name System (DNS) management and root server system management. ICANN is an institution, which not only operates the technical infrastructure of Internet, but also produces policies that relate to national sovereignty issues (Bajaj 2014, 583). The US government is in a position to influence ICANN through the IANA functions contract between the National Telecommunications and Information Administration (NTIA) and ICANN (Kruger 2015; Jayawardane, Larik & Jackson 2015, 6). The paradox here is that although the US is a strong supporter of multi-stakeholderism and opposes an increasing role of governments in governing cyberspace, at the same time, Washington maintains its authority over Internet via the IANA contract (Kruger 2015, 17). Even prior to the Edward Snowden's disclosures about the NSA's surveillance programs, many states were critical of the control that the US exerted over ICANN.

In March 2014, the NTIA announced the intention to transition its stewardship role and procedural authority over key Internet domain name functions to the global Internet multi-stakeholder community. To accomplish this transition, NTIA has asked ICANN to convene interested global Internet stakeholders to develop a transition proposal. Internet stakeholders are currently engaged in a process to develop such a proposal (Kruger 2015, 3). Supporters of this transition argue that such a development will eventually lead to the democratization of global internet governance and strengthen the multi-stakeholder governance. On the other hand, it is still unclear who will replace the US in terms of control. Will ICANN become more accountable or will other states take advantage of the Snowden's revelations and impose a more intergovernmental form of governance? (Kruger 2015, 17-18).

The IGF[9], functions under the aegis of the UN. It was created in 2006 by the World Summit of the Information Society (WSIS) and it aims to bring together various stakeholders in discussions on public policy issues relating to the Internet. The IGF serves as a grassroots discussion forum, where all participants can address the international community; identify emerging issues regarding the management of Internet and shape decisions that will be taken in other forums. The IGF is an open forum that allows developing countries the opportunity to engage in the debate on Internet governance. The obvious disadvantage of the IGF is that it lacks a decision making mandate and the authority to establish policies and regulations. The IGF is therefore useful as a flexible forum for discussions and norms development, but it is questionable whether it has the power to influence the policy making process (Jayawardane, Larik & Jackson 2015, 7).

The NETmundial[10] is another example of multi-stakeholderism that endorses a bottom-up approach regarding Internet governance. In light of the Snowden revelations, the Brazilian government initiated the NETmundial meeting in April 2014. It brought together hundreds of stakeholders from almost 100 countries. The stakeholders represented governments, the private sector, civil society and the academic community (Jayawardane, Larik & Jackson 2015, 7). The NETmundial Multistakeholder statement embraced democratic multistakeholder processes and net neutrality, but failed to reach a complete consensus. Russia and India opposed the NETmundial statement and China and South Africa did not react enthusiastically (Kurbalija 2014, 183).

4. Cyberspace governance in pieces

Cyberspace is not immune to politics. The popular belief that cyberspace can be a libertarian utopia, seems utterly unrealistic in the era of IoT and Big Data. Cyberspace is not the apolitical zone of non-state actors. On the contrary, it is a domain where states seek to exercise their sovereignty. As a result cyberspace governance resembles a power politics game. The case of the ITU and to a lesser extent the recent example of the SCO, colorfully demonstrate that state sovereignty in cyberspace is not only expected, but in some cases it is regarded as the only suitable source of authority (Cornish 2015, 160). The issue at stake is not whether states assert sovereignty *in* cyberspace, but how not to assert sovereignty *over* cyberspace (Slack 2016, 74).

Another point to consider is the future demographic trends in cyberspace. In the past cyberspace was western-dominated, but that will not be the case in the near future. Nowadays, only 30% of world population has access to Internet. The next billions of cyberspace users will originate from the emerging economies that will bring with

[8] For more details see https://www.icann.org/.
[9] For more details see http://www.intgovforum.org/.
[10] For more details see http://netmundial.br/.

them their own values and norms (Deibert 2013, 9). The center of gravity in cyberspace will shift to the east and the south (Nocetti 2015, 111). Many of these states embrace a Westphalian understanding of state that traditionally favors the multilateral governance model.

On the other hand, the analysis of multi-stakeholderism, revealed its shortcomings. There are legitimate concerns regarding the unbalanced representation of civil society groups and private companies, the role they might serve and their ability to actually influence decision making (Dilipraj 2014, 4; West 2014, 9). Multi-stakeholderism does not always lead to a wider range of views or to a more global representation of interests (Pohle 2015). This is not to devalue the role of multi-stakeholderism, but to place this approach into a pragmatic context. After all, the need to create networks of experts, governmental and non-governmental, technical and policy-oriented, should not be underestimated (Slack 2016, 73).

The present state of cyberspace governance commands us to be realistic. While the demand for governance is great, the prospect of an overarching cyber-treaty does not seem feasible, at least in the near future. Treaties are a popular way for concluding agreements between states, but they are not the only way for effective governance. The recent evidence from cyberspace governance seems to confirm Patrick's approach on *minilateralism.* He argues that effective methods of governance occur less in formal institutions and more on regional organizations among like-minded states (2014). The idea behind minilateralism is that it is hard to get 200 states to agree on a single topic. The evidence from climate change, nuclear proliferation, terrorism and trade protectionism demonstrates that it is very hard to achieve international consensus. Multilateralism should not serve as a panacea. Instead, it is better to focus on a smaller number of (like-minded) states, where it will be easier to reach a compromise and generate consensus. States that are not invited in a minilateral forum will denounce this approach as exclusionary, but once agreements are reached and real solutions are generated, minilateral deals will be open to any state that is willing to embrace the values of the original group. It is important to understand that minilateralism does not undermine multilateralism, but instead complements it.

The argument is that since no UN treaty could regulate the whole range of cyber-related issues (cyberwar, cybercrime, protection of civil rights, etc), a more practical approach would be to focus on certain aspects, whether that is the cyber arms race or the protection of intellectual property rights, and develop for each issue norms of behavior and confidence building measures, across a variety of fora. This approach, labeled as *global governance in pieces*, could serve cyberspace in the present transitional phase (Patrick 2014). Instead of relying on large-scale multilateral negotiations that will end in stalemate, it is far more productive to invest in minilateral fora that will deal successfully with a single piece of cyberspace governance every time.

5. Conclusion

The above survey of cyberspace governance illustrated the complexity of the issue. Establishing a social contract for cyberspace, that would involve governments, companies and civil society actors, is easier said than done. The available institutions and the existing body of international law provide adequate tools to regulate a wide range of state activities in cyberspace. Cyberspace governance is still under construction. Cyberspace lacks a single forum or international organization that is responsible for regulating its activities. As a result, governance is spread throughout technical standard setting fora, private sector organizations, civil society groups, states and international organizations. Governance ranges from developing norms and codes of conduct, to signing regional treaties and imposing regulations. Cyberspace governance has reached neither the level of an overarching international treaty, nor that of complete fragmentation.

An effective governance regime requires a synthesis of cyber norms development and confidence building measures among the various stakeholders. This could build a momentum that would gradually lead to bilateral arrangements and multilateral treaties between like-minded states. Cyberspace requires treaties that will develop certain security and human rights standards, and implement mechanisms that will monitor them. In the absence of effective and reliable global governance mechanisms, national institutions remain the only powerful mechanism to implement such standards.

Acknowledgements

The publication of this paper has been partly supported by the University of Piraeus Research Center.

References

Bajaj, K. (2014). Cyberspace: post Snowden. *Strategic Analysis,* 38(4), 582-587.

Barlow, J.P. (1996). A Declaration of the Independence of Cyberspace, last access 20 February 2016, https://www.eff.org/cyberspace-independence.

Betz, D. & Stevens, T. (2011). Cyberspace and the State. Toward a Strategy for Cyber-Power. Adelphi Paper 424. Oxon: IISS & Routledge.

Boothby, W.H. (2014). Conflict Law. Hague: T.M.C Asser Press.

Chertoff, M. & Simon, T. (2015). The Impact of the Dark Web on Internet Governance and Cyber Security. The Centre for International Governance; Global Commission on Internet Governance: Paper Series No.6, last access 20 February 2016, https://www.cigionline.org/sites/default/files/gcig_paper_no6.pdf

Choucri, N. (2012). Cyberpolitics in International Relations. Cambridge, MA: The MIT Press.

Cornish, P. (2015). Governing Cyberspace through Constructive Ambiguity. *Survival,* 57(3), 153-176.

Cukier, K. & Mayer-Schoenberger, V. (2013). The Rise of Big Data. *Foreign Affairs,* 92(3), 27-40.

Deibert, R. & Crete-Nishihata, M. (2012). Global Governance and the Spread of Cyberspace Controls. *Global Governance*, 18 (3), 339-361.

Deibert. R. (2013). Bounding Cyber Power: Escalation and Restrain in Global Cyberspace. The Centre for International Governance Innovation; Internet Governance Papers: Paper no.6, last access 20 February 2016, https://www.cigionline.org/sites/default/files/no6_2.pdf

Deibert. R. (2015). The Geopolitics of Cyberspace after Snowden. *Current History*, 114(768), 9-15.

Demchak, C. & Dombrowski, P. (2011). Rise of a Cybered Westphalian Age. *Strategic Studies Quarterly*, 5(1), 32-61.

DeNardis, L. (2014). The Global War for Internet Governance. New Haven: Yale University Press.

Dilipraj, E. (2014). Internet Governance: The shift from monopoly to multi-party. *National Defence and Aerospace Power*, Issue Brief 99/14.

Dingwerth, K. (2008). From International Politics to Global Governance? The case of nature conservation. *Garnet Working Paper* No.46(8), Institute for Intercultural and International Studies, University of Bremen, last access 20 February 2016, http://www2.warwick.ac.uk/fac/soc/pais/research/researchcentres/csgr/garnet/workingpapers/4608.pdf.

Glen, C. (2014). Internet Governance: Territorializing Cyberspace? *Politics & Policy*, 42(5), 635-657.

Gourley, S.K., (2014). Cyber Sovereignty, in Yannakogeorgos, P.A. & Lowther, A.B. eds. Conflict and Cooperation in Cyberspace: The Challenge to National Security. New York: Taylor & Francis.

Jayawardane, S., Larik, J., & Jackson E. (2015). Cyber Governance: Challenges, Solutions and Lessons for Effective Global Governance. Policy Brief 17. The Hague Institute for Global Justice, last access 20 February 2016, http://www.thehagueinstituteforglobaljustice.org/wp-content/uploads/2015/12/PB17-Cyber-Governance.pdf

Kremer, J.F., & Müller, B. eds. (2014). Cyberspace and International Relations. Theory, Prospects and Challenges. Heidelberg: Springer.

Kruger, L (2015). The Future of Internet Governance: Should the United States Relinquish its Authority over ICANN?. Congressional Research Service Report, last access 20 February 2016, https://www.fas.org/sgp/crs/misc/R44022.pdf.

Kurbalija, J. (2014). An Introduction to Internet Governance. Msida: DiploFoundation.

Lewis, J.A., (2010). Cybersecurity: Next Steps to Protect Critical Infrastructure, testimony to the US Senate Committee on Commerce, Science and Transportation, 23 February 2010, last access 20 February 2016, https://www.gpo.gov/fdsys/pkg/CHRG-111shrg57888/pdf/CHRG-111shrg57888.pdf.

Mihr, A. (2014). Good Cyber Governance: The Human Rights and Multi-stakeholder Approach. *Georgetown Journal of International Affairs, International Engagement on Cyber IV,* 24-34, last access 20 February 2016, http://schd.ws/hosted_files/igf2015/6b/Good%20CyberGovernance-Mihr-2014.pdf.

Nocetti, J. (2015). Contest and conquest: Russia and global internet governance. *International Affairs*, 99(1), 111-130.

Nye, J. S. (2014). The Regime Complex for Managing Global Cyber Activities. The Centre for International Governance; Global Commission on Internet Governance: Paper Series No. 1. https://www.cigionline.org/sites/default/files/gcig_paper_no1.pdf

Nye, J.S. & Donahue, J. (2010). Governance in a Globalizing World. Washington DC: Brookings Institution Press.

Patrick, S. (2014). The Unruled World. The Case for Good Enough Global Governance. *Foreign Affairs,* 93(1), 58-73.

Pohle, J. (2015). Multistakeholderism unmasked: How the NetMundial Initiative shifts battlegrounds in internet governance. Global Policy, last access 20 February 2016, http://www.globalpolicyjournal.com/blog/05/01/2015/multistakeholderism-unmasked-how-netmundial-initiative-shifts-battlegrounds-internet.

Rosenau J. & Czempiel, E.O. eds. (1992). Governance without Government: Order and Change in the World Politics, Cambridge, UK: Cambridge University Press.

Rosenau, J. (1995). Governance in the Twenty-First Century. *Global Governance,* 1 (1), 13-43.

Slack, C. (2016). Wired yet Disconnected: The Governance of International Cyber Relations. *Global Policy,* 7(1), 69-78.

Tabansky, L. (2011). Basic Concepts in Cyber Warfare., *Military and Strategic Affairs,* 3(1), 75-92.

Weber, R.H. (2013). Internet of Things - Governance quo vadis?. *Computer Law & Security Review*, 29(4), 341-347.

West, S. (2014). Globalizing Internet Governance: Negotiating Cyberspace Agreements in the Post-Snowden Era. Conference Paper, *TPRC 42: The 42nd Research Conference on Communication, Information and Internet Policy*, last access 20 February 2016, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2418762##.

Privacy Concerns of TPM 2.0

Ijlal Loutfi and Audun Jøsang
University of Oslo, Norway
ijlall@ifi.uio.no
josang@ifi.uio.no

Abstract: The goal of trusted computing is to provide solutions that allow users to bootstrap trust into their machines based on hardware. The flagship technology for trusted computing is the Trusted Platform Module (TPM) which is specified by the Trusted Computing Group (TCG). TPM hardware chips are currently embedded in 1 billion devices. One of the services that sets TPMs apart from other hardware-enabled security technologies is integrity attestation. However, integrity attestation has been criticized for allowing third party entities that use remote attestation to breach end user's privacy. The recent TPM 2.0 specification has as one of its goals to alleviate this issue. In this paper, we showcase how the new definition of privacy and its corresponding solutions have weaknesses. Solutions which are aimed at protecting end users from third-party privacy attacks have the paradoxical side-effect of exposing end users to potential tracking by manufacturers and other law enforcement entities. We propose two solutions to mitigate this issue.

Keywords: trusted computing, TPM 2.0, remote attestation, endorsement key, tracking, privacy

1. Introduction

In the 1990s, it became increasingly obvious to people in the computer industry that the Internet was going to revolutionize the way personal computers were used, and that commerce was going to move toward this environment. This immediately led to a realization that there was a need for increased security in personal computers (Challener, 2015). This realization has been strengthened by the ubiquitous use of computing devices, as well as by the sensitivity of the business processes that are moving online, such as health, education, and research. However, this goal was challenged by the early hardware and security design principles, which did not take into account security requirements. Hence, a new hardware-based standardized security solution became imperative. The goal of such a solution would be to become an anchor on which new architectures can be built (Challener, 2015).

In 2003, a proposed security solution came in the form of a chip called the Trusted Computing Platform, TPM. Its specifications were released and are maintained by the Trusted Computing Group (TCG). Now, the TPM is present in almost all of computing devices worldwide, and its most recent specification, TPM 2.0, was released in 2014 (TCG-Architecture, 2014). TPM services include all of the security mechanisms traditionally available in other hardware-enabled security solutions, such as providing secure storage for cryptographic material and acting as a crypto processor. In addition, the TPM differentiates itself by its remote integrity attestation capability. Remote attestation allows a platform to cryptographically attest its state to a third party. The latter can then decide on whether it wants to pursue a transaction with the TPM-equipped platform, based on its evaluation of the provided attestation state (TCG-Part1, 2014).

While remote attestation offers, theoretically, significant security advantages, its implementation and practical application has not taken off as expected (Lyle, 2009). The main criticism against remote attestation has focused on 2 aspects: the impracticality of the TPM infrastructure management, and the breach of end users' privacy. The TPM 2.0 specifications released in April 2015 came to alleviate the management and privacy concerns. While we believe that the new specification has succussed in mitigating the management issues, it has increased the risk of privacy breaches. This is due to the way they approach and define privacy. Indeed, the used threat model does not include TPM manufacturers and law enforcement entities as potential threats to end user privacy, which leaves the door open to tracking applications to be built on top of the TPM infrastructure. In a post-Snowden world, assuming the existence of such applications is not a far-fetched thought. The focus of this paper will be on analysing the TPM specifications, and showcasing how tracking can be achieved in such an environment.

We start this paper with a description of the trusted computing paradigm and its flagship technology, TPM. In Section 3 we take a deep dive into the newly released TPM 2.0 specifications. Section 4 presents an overview of the available literature that discusses the privacy issues of TPM. In Section 5, we discuss how the privacy

definition introduced in TPM 2.0 is incomplete, and how it can enable tracking of end users. In Section 6, we propose mitigation solutions. We close with a discussion section.

2. Trusted computing and the trusted platform module

The concept of the Trusted Computing was developed by an industry consortium, referred to as the Trusted Computing Group. It aims at protecting computing infrastructure and billions of end points, based on a hardware root of trust (TCG-Architecture, 2014). The principal mechanism for achieving this goal is to verify and enforce known, and thereby trusted, configurations of computing platforms. The verification of platform configuration rests on establishing a complete chain of trust, i.e. a verified list of all hardware and software that has been installed on a platform (Lyle, 2009)(Llopart, 2013). This chain of software can then be compared to a list of known 'trusted' software modules, in contrast to traditional approaches such as virus scanners that try to recognize and eliminate instances of bad software in isolation from the rest of software modules on a computing platform (Dietrich, 2012).

The TPM (Trusted Platform Module) is of one of the main building blocks of this paradigm. TPM is defined by the TCG as a computer chip micro-controller that is attached to the motherboard. The technologies proposed by the TCG are centred on the TPM. In a basic server implementation, the TPM is a chip connected to the CPU (Lyle, 2009). The main services offered by TPMs are: Secure Storage; Integrity Measurement; and Remote Integrity Attestation. The latter is the focus of the paper, and its details are explained in Section 3 (TCG-P2. 2014).

3. Integrity reporting and attestation

Remote integrity attestation works on the principle that a platform can be trusted if all the software and hardware it has run (its `configuration') can be identified and verified by a relying party (Lyle, 2009)(Martin, 2008). Section 3.1 explains how the TPM enables the collection of measurements about the designated software and hardware, and Section 3.2 focuses on how these measurements can be communicated to a third party through the process of remote integrity attestation.

3.1 Integrity measurement

The TPM can securely store artefacts (encryption keys, passwords, certificates). It can also store platform integrity measurements that are used to verify the platform integrity. In the TPM jargon, this is expressed by saying that it helps to ensure that the platform remains trustworthy. This is possible thanks to the 16 PCRs (Platform Configuration Registers) it provides. A PCR is a 256/512 bit wide register that can hold a hash digest. Each digest corresponds to a measurement of a piece of software or hardware present on the platform (TCG-P2. 2014). It is not possible to write directly to a PCR. The only allowed PCR operation is **extend(x)**: This operation calculates the new value of a PCR as a hash digest of the concatenation of the old value and the new value **x** (TCG-P3. 2014) which then becomes a hash chain. By definition, the extend operation is non-commutative, meaning that the tracking of the order of events is guaranteed. Also, because the size of the hash digest is fixed, a PCR has the capacity to store an arbitrary number of measurements.

3.2 Remote integrity attestation

Attestation is a more advanced use case for PCRs. In a non-TPM platform, remote software can not usually determine a platform's software state. If the state is reported through strictly software means, compromised software can simply lie to the remote party. A TPM attestation can theoretically offer cryptographic proof of software configuration state. This state is communicated through a set of digitally signed PCR measurements. Indeed, the attestation is a TPM quote: a number of PCRs are hashed, and that hash is signed by a TPM key. If the remote party can validate that the signing key came from an authentic TPM, it can be assured that the PCR digest report is authentic and has not been altered (Challener, 2015).

Figure 1 illustrates the principle of remote attestation based on the TPM. The entities in the left-hand side reside in the client platform, and the challenger on the right-hand side represents the remote party.

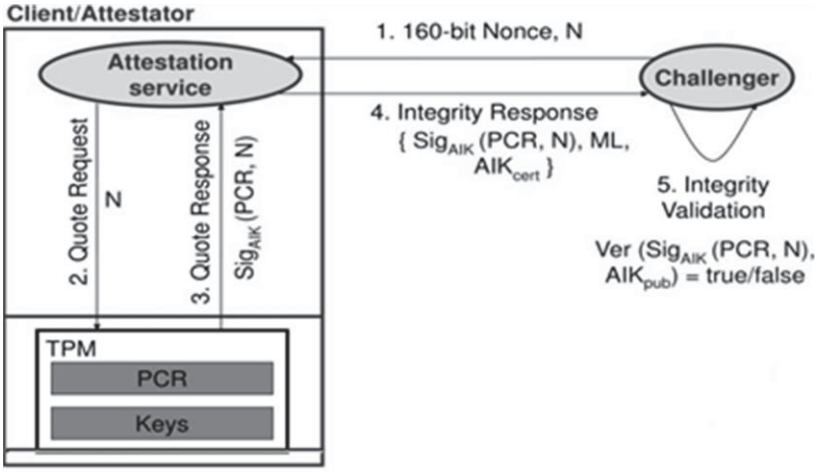


Figure 1: Remote attestation message exchange (Lavina, 2010)

3.3 Issues around remote integrity attestation

In this section, we present a concise summary of the TPM privacy issues that our survey of the literature has revealed.

First of all, enabling integrity measurement and reporting would allow a remote party to have information about what software is being run on the user's platform. This can be problematic in an open ecosystem where a remote party can decide to take different actions based on what software is run by the platform it is interacting with, possibly based on unfair discriminatory criteria (Sadegui, 2004) (Kühn, 2007).

Furthermore, if a remote party has malicious intentions, it can identify the specific software modules running on the platform, and target its attack based on the known vulnerabilities of those software modules**.**

We concluded that most of the privacy issues discussed boil down to two main questions:

- How can TPM provide a means to prove that a key was created and was protected by a TPM without the recipient of that proof knowing which TPM was the creator and protector of the key? (Challener, 2015)
- How can we ensure different recipients of remote attestation are not going to link the identities of the users, and augment their knowledge about him and his platform, to more than they were supposed to have?

TCG has tried to address these concerns first in TPM 1.2 and then in TPM 2.0. The first draft specification of TPM 1.2 proposed the use of separate privacy CA. However, privacy groups complained about the difficulty of implementing and operating privacy CAs (Challener, 2015). In TPM 1.2, this led to the inclusion of new commands for a second method of anonymizing keys to help address this concern — direct anonymous attestation (DAA) — which is based on group signatures and provides a relatively complicated method for proving that a key was created by a TPM without providing information as to which specific TPM created it. The advantage of this protocol is that it lets the AIK (Attestation Identity Key) creator choose a variable amount of knowledge they want the privacy CA to have, ranging from perfect anonymity (when a certificate is created, the privacy CA is given proof that an AIK belongs to a TPM, but not which one in particular), to perfect knowledge (the privacy CA knows which EK (Endorsement Key) is associated with an AIK when it provides a pseudonymous certificate for the AIK). The difference between the two methods is apparent when TPM hardware is broken into and inspected, whereby a particular EK's private key is leaked, and potentially published in the Internet. At this point, a privacy CA can revoke the certificate if it knows that it created the privacy certificate associated with that particular EK, but cannot revoke the certificate if it does not know the association between certificate and EK. Once more, DAA failed to be adopted in the field (Challener, 2015).

In TPM 2.0, the TCG responded to the above mentioned privacy issues and the criticism against the security solutions of TPM 1.2, by creating 3 separate key hierarchies. This seems to have tamed down those voices of criticism. We believe that this architecture will alleviate the management burden. However, it will exacerbate the privacy concerns.

The privacy concerns of TPM 2.0 are due to ***the way privacy is defined by TCG***. In the specifications of trust requirements for TPM 2.0, TCG excludes the manufacturers of TPM chips and computing platforms from the set of potential privacy threats. This assumption is unrealistic for corporate and private users of computing platforms, especially in the post-Snowdon world we live in, where we know that (secret) state sponsored tracking and mass surveillance is a reality. The next section extends on the justification for our claim.

4. Incomplete privacy trust model

The previous sections presented the privacy issues raised by the community, as well as the TCG's response in terms of the TPM 2.0 specifications. We believe the trust model upon which privacy is defined for TPM 2.0 is incomplete. We will examine this claim by examining how TCG defines privacy.

> *The inability of remote parties receiving TPM digital signatures to correlate them*
>
> *—to cryptographically prove that they came from the same TPM. A user can use different signing keys for different applications to make correlation difficult. The attacker's task is to trace these multiple keys back to a single user (Challener, 2015)*

This definition models remote transaction parties as the sole potential threat to end users' privacy. It remains silent about potential threats from TPM manufacturers and law enforcement entities. Hence, by omission, the latter players are assumed to be entities that the end user assumes will not try to breach his/her privacy. That this is an unrealistic assumption is intuitively clear when considering the Snowden revelations (Greenwald *et al.*, 2013) (Menn, 2013). That is, a law enforcement entity can and does subpoena different commercial manufacturers, and legally coerces them to hand in private end users' data.

As an analogy, consider the manufacturer of self-encrypting drives (SEDs), i.e. data storage devices that automatically encrypt and decrypt data that is being written to, and read from the device, under a key defined by the owner. Assume now that the same threat model as for TPM 2.0 is being used, which would imply that the SED manufacturer can read data on all SEDs it has produces and sold, because it has a way to recover encryption keys defined by owners. This type of SEDs would certainly not succeed in the market if the key escrow capability were known. No users in their right mind would buy such SEDs.

For TPM 2.0, the market situation is different, because users do not actually buy TPM chips. Instead, TPM chips come bundled with the computer platform that users buy, typically without their knowledge.

The TCG aims at making the TPM a cornerstone of the TCB (Trusted Computing Base) of computing platforms, where privacy is assumed to be one of its main functions. From that perspective it is reasonable to expect that the TPM genuinely supports the privacy of end users in a transparent fashion. It is therefore surprising to read in the TPM 2.0 specifications that end-user privacy has been partially traded off to give TPM manufacturers the power to identify and trace end-user computing platforms.

4.1 TPM privacy compromising design decisions

Because of this incomplete definition of the privacy trust model, the TPM 2.0 specifications define insufficient and inadequate mechanisms to protect end users from law enforcement entities and the TPM manufacturers breaching their privacy. Two of these problematic design decisions are presented in the following sections:

4.1.1 New hierarchies: The always-on platform hierarchy

A hierarchy is a collection of entities that are related and managed as a group. Those entities include permanent objects (the hierarchy handles), primary objects at the root of a tree, and other objects such as keys in the tree.

TPM 1.2 had only one hierarchy, represented by the owner authorization and storage root key (SRK). There can be only one SRK, always a storage key, which is the lone parent at the top of this single hierarchy. The SRK is generated randomly and can not be reproduced once it has been erased. It can not be swapped out of the TPM. Child keys can only be created and wrapped with (encrypted by) the SRK, and these child keys may in turn be storage keys with children of their own. However, the key hierarchy is under the control of the one owner authorization; so, ultimately, TPM 1.2 has only one administrator.

The TPM 1.2 hierarchy architecture led to considerable challenges to how TPM platforms would be managed. This is mainly because of the overlapping authorization domains for the TPM firmware, TPM owner (e.g.: IT administrator who own the platform in which the TPM is embedded), and TPM end user. In order to overcome the management challenges, TPM 2.0 introduced a new architecture, which we believe has increased the risk of tracking TPM end users (Challener, 2015). TPM 2.0 defines four different key hierarchies:

- ***Standard storage hierarchy:*** Replicates the TPM 1.2 family SRK for the most part
- **Platform hierarchy:** Used by the BIOS and System Management Mode (SMM), *not* by the end user
- **Endorsement hierarchy or privacy hierarchy**: Prevents someone from using the TPM for attestation without the approval of the device's owner
- ***Null hierarchy:*** Uses the TPM as a cryptographic coprocessor

The platform hierarchy, which is the focus of this section, is intended to be under the control of the platform manufacturer, represented by the early boot code shipped with the platform. The platform hierarchy is new for TPM 2.0. In TPM 1.2, the platform firmware could not be assured that the TPM was enabled. Thus, platform firmware developers could not include tasks that relied on the TPM (TCG-Architecture, 2014).

As the most privileged hierarchy, the Platform Hierarchy *is enabled* at reboot. The intent is that the platform firmware will generate a strong platform authorization value (and optionally install its policy). Unlike the other hierarchies, which may have a human enter an authorization value, the platform authorization is entered by the platform firmware. Therefore, there is no reason to have the authorization persist (and to find a secure place to store it) rather than regenerate it each time (Challener, 2015).

The problematic part of the platform hierarchy in our opinion is that:

- It has its own enable flag
- The platform firmware decides when to enable or disable the hierarchy. While the TCG intent of this design decision is for the TPM to always be enabled and available for use by the platform firmware and the operating system, this decision clearly takes consent away from end users.
- The platform hierarchy also does not need to use the same cryptographic material that is used in the rest of the TPM, making it ever more obscure.

4.1.2 Re-Certification

Every hierarchy has a list of keys and object that belong to its control domain. The keys and data under any given hierarchy are encrypted using the primary key of that same hierarchy. In TPM 1.2, the Endorsement Key sitting at the top of the TPM hierarchy can never be erased once it has been embedded in the TPM chip by its manufacturer. This was a clear breach of end user's privacy, as all subsequent keys generated by the TPM could be traced back to a single EK, and hence, identifies the platform. The proposed solutions (DAA and privacy CA) are not real solutions for this privacy issue, as they would still allow the TPM manufacturers and low enforcement entities to track end users. Furthermore, these solutions were never really implemented in TPM 1.2, leaving end users exposed also to third party attestators privacy attacks.

In TPM 2.0, each one of the 4 new hierarchies could have a different seed from which all subsequent hierarchy keys are generate. Furthermore, these seeds could be deleted even after the TPM has been shipped to end users. At first, this design decision sounds like an appropriate solution to guard against privacy threats from TPM manufacturers and law enforcement agencies. However, it is very surprising that TPM 2.0 has introduced an extra mechanism, called *re-certification*, that always maintains a cryptographically traceable link from the platform Hierarchy controlled by the TPM manufacturer, to the other hierarchies, even when the root keys of the other hierarchies are deleted and re-generated by the user. This mechanism gives TPM manufacturers the power to trace TPM-equipped platforms no matter what. Ironically, in the jargon of the TPM 2.0 specifications, this is described as *"to maintain the chain of trust"*.

Indeed, in the platform hierarchy, the primary seed (EPS) can be replaced by a new value. However,

> *"the TPM specification allows cross certification of keys between the Platform hierarchy and the Endorsement hierarchy under control of the platform firmware. Cross certification allows a chain of trust to be maintained as the seeds are changed (Challener, 2015). "*

Figure 1 illustrates the ownership of key hierarchies before and after re-generation by the end user. Also shown is how the TPM manufacturer can keep track of key hierarchies it does not own.

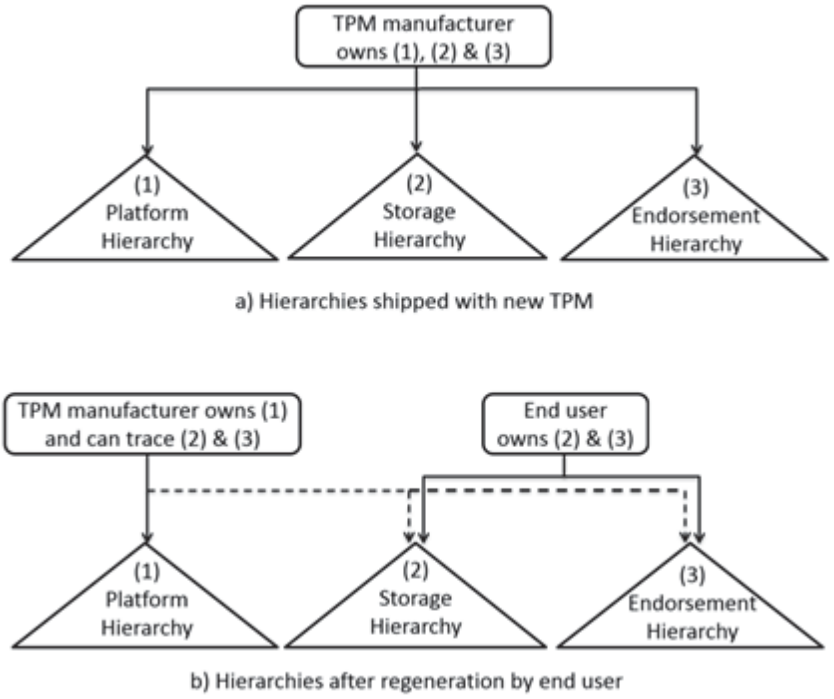


Figure 2: Re-certification of key hierarchies in TPM 2.0

5. Proposed solutions

Any reasonable solution to the above discussed privacy issues would of course require a radical change of the design of TPM, as it would question the fundamental definition of privacy that the specifications are built upon. While we do wish this to be the case, we do acknowledge the impracticality, or rather the difficulty of pushing for such a change. Hence, in this paper we are advocating for two main measures in order to mitigate the risks of mass tracking.

5.1 Elimination of re-certification

If TPM is to truly succeed and thrive in an enterprise environment, and/or be adopted in a serious manner for large scale national deployment; and if it is to become an attractive architecture for developers to build their security upon, the TPM should allow for a way for its keys and cryptographic data to be completely outside of the PKI of the TPM manufacturers. While this obviously would require more infrastructure management from the IT organization that would choose such a design, it would mean that they can be assured that their data will be completely under their sovereignty, and that their privacy-sensitive data has no way of being compromised from manufacturers. This requirement is ever more important in today's geopolitical situation, in which state spying on each other is common currency. In this post-Snowden era, one can not assume that a company in a foreign nation state will not hand in private data about end users, should the authorities of that nation state request it. This proposed solution, if adopted, would be in line with the TPM decision to allow each manufacturer to embed cryptographic algorithms that they trust. These two design requirements together are what could make TPM truly a device that not only defends against malware, but also against mass tracking by potential adversaries. TPM would in this case be the end user's true advocate.

5.2 Disclosure

In case of TPM 2.0 where the TPM has a platform hierarchy over which the owner has no control, and which can be enabled whether the owner of the platform desires it or not, we believe that the updates and any other

possible cryptographic communication that goes into and out of it should be made visible to end users, whenever they are interested. If they are not able to disable it completely, it would be more ethical for them to be aware of what is happening in their own platform. Hence, we suggest that a monitoring application should be provided to end users in order to notify them about passive actions that do not require their active participation, and yet are happening. This could allow consent from end users in order to perform TPM platform updates for instance. Such an application is technically possible, so it is really a matter of willingness whether we would see it happening or not.

6. Conclusions

In this paper, we have discussed the shortcomings of the privacy definition provided by the Trusted Computing Group, and around which TPM 2.0 specifications have been designed. We conclude that TPM 2.0 does indeed strengthen the protection of end users from third party applications that use the remote attestation service. However, it does not protect end users from tracing and tracking by platform and TPM manufacturers, nor from law enforcement authorities that can subpoena them to hand in user sensitive data. In this post-Snowden era, it is not far-fetched at all to talk about mass surveillance at a state scale. This problem becomes ever more serious when considered in the context of states interfering with the sensitive personal data of foreign end users, just because the platform and TPM manufactures happen to be in their legal jurisdiction. This is no light matter to play around with, given the type and amount of sensitive information remote attestation by the TPM can reveal about its platform owner. We have proposed two solutions aimed at mitigating the risk of tracing and tracking. We believe that these issues should be highlighted, not only for ethical reasons, but also because it will give the TPM a real chance in being adopted in the enterprise market. If so, it would be a powerful tool to add to our security defence arsenal in a cyber-landscape of increasing threats.

References

Dietrich, Kurt *et al*. (2007). A Practical Approach for Establishing Trust Relationships between Remote Platforms using Trusted Computing. In *Trustworthy Global Computing*, LNCS 4912, pp 156-168. Springer 2007.

Greenwald, Glenn and MacAskill, Ewen (2013). *NSA Prism program taps in to user data of Apple, Google and others*, The Guardian, 7 June 2013, http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data

Jiewen, Yao and Zimmer, Vincent (2014). *A Tour Beyond BIOS with the UEFI TPM2 Support in EDKII*. Intel White paper, 2014. https://firmware.intel.com/sites/default/files/resources/A_Tour_Beyond_BIOS_Implementing_TPM2_Support_in_EDKII.pdf

Kühn, Ulrich and Selhorst, Marcel and Stüble, Christian (2007). Realizing property based attestation and sealing with commonly available hard- and software. Proceedings of the 2007 ACM workshop on Scalable trusted computing (STC'07), pp.50-57. ACM New York, 2007.

Lavina, Jain and Vyas, Jayesh (2010). *Security Analysis of Remote Attestation.* CS259 Project Report. Stanford University, 2010.

Lyle, John and Martin, Andrew (2010). Trusted computing and provenance: Better together. In Proceedings of the *2nd Conference on Theory and Practice of Provenance (TAPP'10)*, pages 1. 2010.

Lyle, John and Martin, Andrew (2009). On the Feasibility of Remote Attestation for Web Services. *International Conference on Computational Science and Engineering (CSE '09)*. IEEE 2009.

Martin, Andrew (2008). *The ten page introduction to trusted computing*. Technical report CS-RR-08-11, University of Oxford, 2008.

Menn, Joseph (2013). *Exclusive: Secret contract tied NSA and security industry pioneer*, Reuters, 20 December 2013, http://www.reuters.com/article/2013/12/20/us-usa-security-rsa-idUSBRE9BJ1C220131220

Pelegri-Llopart, Eduardo and Yoshida, Yutaka and Moussine-Pouchkine , Alexis (2007). *The GlassFish Community Delivering a Java EE Application Server*, Sun Microsystems, 2007. https://glassfish.dev.java.net/faq/v2/GlassFishOverview.pdf.

Schneier, Bruce (2015). *Everyone wants to have security, but not from them*. https://www.schneier.com/blog/archives/2015/02/everyone_wants_.html.

Sadeghi, Ahmed-Reza and Stuble, Christian (2004). Property-based attestation for computing platforms: caring about properties, not mechanisms. *Proceedings of the 2004 Workshop on New Security Paradigms (NSPW '04)*. 2004.

TCG. TPM Main Part 1, Design Principles (2014). Technical report, Trusted Computing Group. 2014.

TCG. TPM Main Part 2, TPM Structures (2014). Technical report, Trusted Computing Group. 2014.

TCG. Trusted Platform Module Summary. Technical report, Trusted Computing Group. 2014.

TCG. TPM specification: Architecture overview. Technical report, Trusted Computing Group. 2014.

TCG. Attestation Identity Key (AIK) Certificate Enrolment Specification: Frequently asked questions. 2011.

TCG. Endorsement Key (EK) and Platform Certificate Enrolment Specification. 2013.

Will, Arthur and Challener, David and Goldman, Kenneth (1015). *A Practical Guide to TPM 2.0*. Apress Open, Springer New York, 2015

Future Digital Forensics in an Advanced Trusted Environment

Markus Maybaum and Jens Toelle
Fraunhofer FKIE, Bonn, Germany
markus.maybaum@fkie.fraunhofer.de
jens.toelle@fkie.fraunhofer.de

Abstract: Automation in digital forensics has been subject to intensive research in various disciplines. One of the main challenges in the field is finding an efficient automated procedure to properly collect persuasive evidence and to archive it in a secure manner. In this paper we discuss this challenge from a technical perspective. We introduce the concept of a Future Advanced Trusted Environment framework and discuss the kind of incidents this framework can identify autonomously and those for which human intelligence is still required. Based on this framework we introduce the Trusted Forensic Module, a new hardware security extension to standard computer systems. We explain this new module in detail and show its advanced concept.

Keywords: trusted computing, digital forensics, future advanced trusted environment

1. Introduction

Cyber-crime is a business, and to be precise cyber-crime is a very profitable business. From the perspective of the monetary damage caused by cyber-crime over the last 15 years in the U.S. alone (see Figure 1), we can reasonably suggest that fighting cyber-crime is increasingly important, especially when taking into account that in the same period the government invested millions of dollars in cyber-crime fighting capabilities. For example, the FBI (FBI2016) established a Cyber Division and trained specially equipped cyber squads which can travel globally to help other countries. Considering the continuing rise of cyber-crime, this did not help to significantly reduce the damage to the U.S. economy. Consequently, Figure 1 not only illustrates the need to fight cyber-crime due to its devastating impact on a state's economy, but at the same time unfortunately shows that the investments made so far are far from being enough.

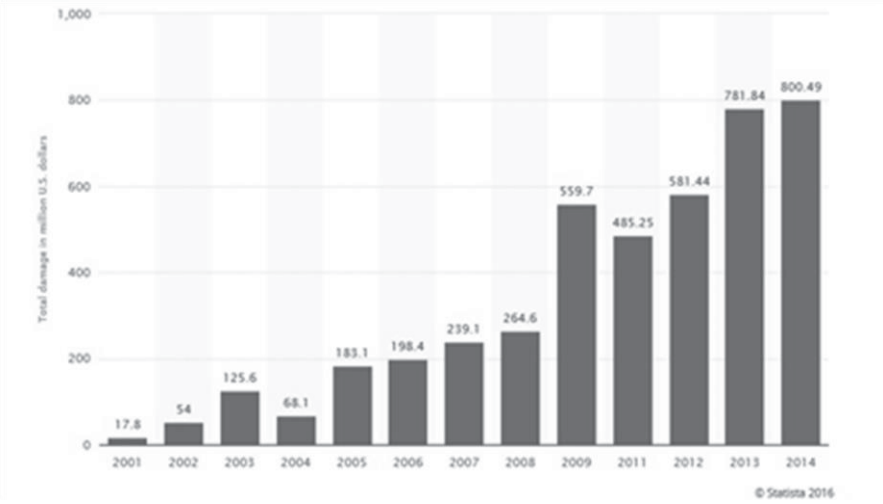


Figure 1: Monetary damage caused by cyber-crime (Statistica 2016)

One of the reasons for this unpleasant situation may be the obstacles for investigators scrutinising the source of a cyber-attack. Even if the source has been properly identified there remains difficulty in collecting evidence, so that the attacker can successfully be prosecuted. Without prosecution, cyber-crime will continue to be a problem. In the light of state-of-the-art digital forensic working patterns, and the need for a transparent chain of custody, it is easy to understand why the gathering of evidence is very costly in terms of time and in terms of the resources required. Experts in the field of digital forensics claimed years ago (Richard, Roussev 2006) that it would be essential in the future to design and establish automated digital forensic systems that work autonomously, with a minimum of interaction with human investigators. It is reasonable to believe that this will apply to both evidence collection and evaluation procedures.

The intent of this paper is to provide clear information about new technical options within digital evidence collection. In particular, we propose a practical framework for automated autonomous digital forensic systems.

We believe that a major deficiency in state-of-the-art forensic techniques is that the chain of custody is vulnerable due to missing trust relationships: trust of human beings for their machines (especially after a security incident) and trust relationships between machine components, which is an essential requirement when talking about automated autonomous digital forensics. We therefore focus our research on trusted environments; we believe, from a technical standpoint, that automated autonomous evidence collection is possible if such trust deficiencies can be overcome. In this paper we present a proof of concept based on the idea of a Future Advanced Trusted Environment, using the Trusted Computing concept.

2. Related work

Digital forensics is a young discipline which still requires intense research efforts in multiple disciplines. In general, the work described in this paper relies on the basic principles of digital forensics as collected by Brenzinski and Killalea (2002) in the Request for Comments 3227 (RFC3227) on best practices in the field of digital forensics, as well as on Casey's (2010) Handbook of Digital Forensics and Investigation and Cohen's (2007) Challenges to Digital Forensic Evidence. We particularly refer back to inspiration from Richard and Roussev (2006), in their article on next-generation digital forensics, who claimed that the future of digital forensics will focus on automation and autonomous evidence collection, already giving early indications that hardware security extensions might be the key to success. Significant work has been done in that field already, and must be mentioned here. Carrier and Grand (2004) designed a PCI card with a hardware extension implementing a procedure for acquiring volatile memory. Their card can copy memory to an external storage device. Wang et al.'s (2011) paper on Firmware-assisted Memory Acquisition discusses the problem of how to reliably and consistently retrieve a volatile machine state without disrupting operations. They suggest enhancing commercial PCI network cards and the current x86 implementation of the system management mode to make them capable of replicating both the volatile memory and critical CPU registers of automated collection devices.

Another source of inspiration, and closer to our work, is a framework presented by Case et al. (2008). They demonstrated a framework for automatic evidence discovery and correlation from a variety of forensic targets. In particular, they demonstrated an open-source memory analysis tool they called "Ramparser", which is capable of performing an automated forensic analysis of a Linux system. Finally, research and development, as well as the standardisation of Trusted Computing (TC), are the main missions of the Trusted Computing Group (TCG). The TCG is supported by numerous international research teams from academia as well as from industry, and governmental organizations also contribute significantly to this project across the globe (Trusted Computing Group 2016). The main focus of TCG's work is currently standardisation of the Trusted Platform Module (TPM) and the integration of new security applications.

3. Future advanced trusted environment – an enabler for autonomous digital forensics

In this section we introduce the concept of what we called the Future Advanced Trusted Environment (FATE). We will demonstrate our proposals using the example of TC and the security extensions that we have already designed for it. TC offers a hardware platform in which the integrity of IT systems is verified using a structured file-based signature hierarchy for all executable system components. Based on the 'chain-of-trust' concept, all system components from the BIOS up to the user applications are protected by hash checksums so that any unauthorised modification can be seen immediately (see Figure 2).

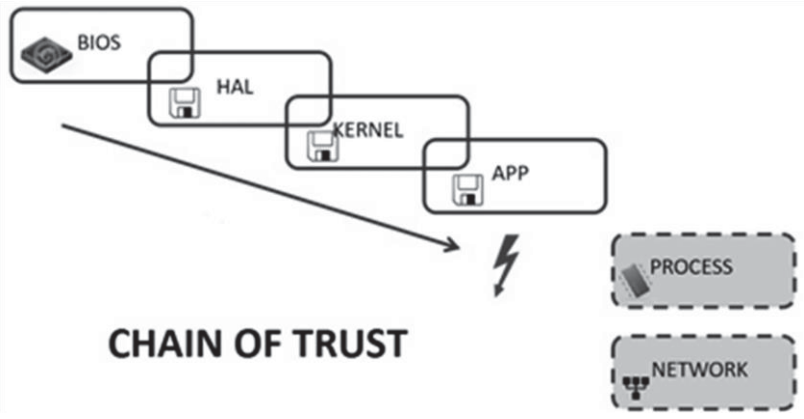


Figure 2: Chain of trust

TC has proven it's resilience against well-known file-based attacks in real life, but flaws still exist at both ends of the chain. The flaw at the front end of the chain involves the vulnerabilities of processes to exploitation techniques when the TC protected file is loaded into the volatile process memory and is no longer monitored by the TPM. The flaw at the back end of the chain is the stand-alone design of the TPM: the detection of a breach of integrity is only reported by the TPM to the operating system, and remains on the compromised machine. The architectural approach of the FATE is to eliminate these flaws by implementing hardware security extensions. For this purpose we designed an Attack Recognition Module (ARM) (Maybaum & Toelle, 2015) as a new module inside the TPM (Figure 3) introducing three new core functionalities: Control Flow Integrity (CFI), an Integrity Supervisor Unit (ISU), and a Peer Communication Unit (PCU). In our design, we also added a fourth module, File and Firmware Monitoring (FFM), where the TPM's existing state-of-the-art functionality (as shown in Figure 2) is now accommodated to improve performance; we call this the Enhanced-TPM (ETPM).

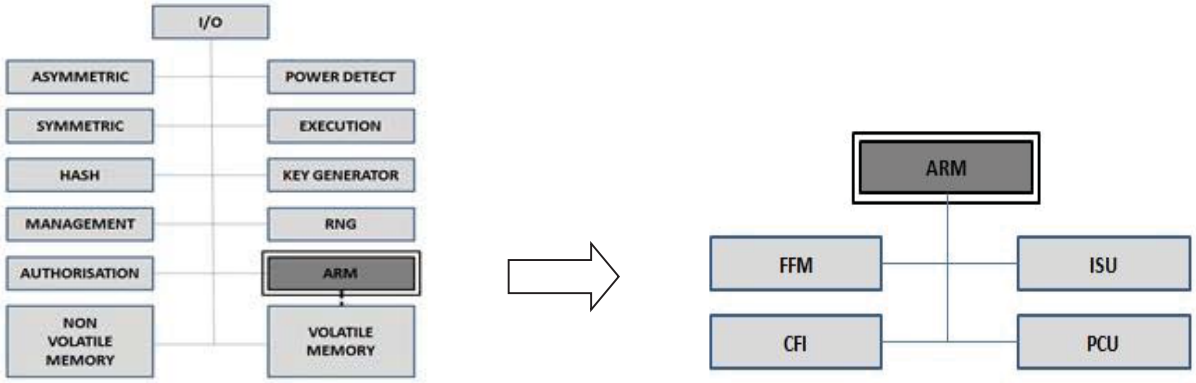


Figure 3: Enhanced-TPM

In this paper we introduce the core concept and basic principles of the FATE, using the example of the ETPM to demonstrate the concept of our proposed technology. We include references to publications describing the technical implementation in more detail: we cannot re-cite all the technical features of the ETPM here, so to fully understand the features of the ARM, please refer to Maybaum and Toelle (2015). Basically, the ISU supervises the integrity of the system and takes action if it detects any integrity breach. The PCU communicates with the peers at the link-layer (Layer 2). If the ISU reports an integrity breach (internally), the PCU notifies the connected peers about this breach. All PCU communication is based on a new secure protocol at the link-layer that we called the Trusted Address Resolution Protocol (TARP), based on secure channels using Diffie-Hellman encryption (Diffie & Hellman 1976). The core functionality of the ETPM is implemented into the CFI module which is able to reliably detect integrity breaches. If a running process is exploited, the CFI will recognise the control flow deviation and notify the ISU to begin incident handling.

3.1 Incident detection and handling in a trust-based environment

The trigger of any digital forensic activity is always the detection of an incident. In this section we want to analyse the kind of incidents that can be identified by the ETPM. This requires a detailed understanding of the CFI module in the ARM. The CFI module implements a concept we have called Trusted Control Flow Integrity (TCFI). TCFI uses the ideas and methods of traditional TC and implements them at process level by using enhanced known CFI safeguards. These safeguards enable the CFI module to reliably recognise any control flow deviation into unauthorised code execution (malware) and respond appropriately to stop the offending action. In Maybaum (2015) we proved the concept of TCFI with examples of two techniques: Trusted Branch Table Control Flow Integrity (TBT-CFI) and Trusted Checkpointing Control Flow Integrity (TCP-CFI). TBT-CFI uses branch tables created at compile time to monitor dynamic branches during process execution. With TCP-CFI, the programme control flow is enriched by a set of check points which again can be verified by known TC concepts. Applying both concepts in combination increases the resilience of the protected system significantly, and allows the system to reliably identify exploitation-based incidents. In addition to unauthorised manipulation of files and firmware, illicit code execution can also be seen in the system's memory and can be made subject to incident handling and digital forensics. So, in principle, the ETPM is capable of identifying incidents based on programme code manipulation. For this work, we therefore define the term 'security incident' explicitly as a control flow integrity breach.

In case of a Security Incident, the process being executed becomes suspicious and the ISU will report an integrity breach status to the operating system. In the current TPM design, this can be implemented by a status result parameter being added to TPM function calls, however, the status can only be monitored and checked asynchronously. In a FATE, this is overcome because the ETPM has a more pro-active role and will be able to send its own interrupts to the OS to stop the processing of the suspicious process. In accordance with the current design of the ETPM, an integrity breach being monitored by the ISU is first of all reported to the peering systems by the PCU so that the potentially compromised system can be isolated (for example, by blacklisting). From an abstract viewpoint, the peering systems build a network of trusted machines called the Trusted Integrity Networks (TIN). The link-layer network between the PCUs builds a trusted segment. This concept has been designed to prevent peering systems from being infected in cases where the integrity breach is caused by malware that spreads through networks. In earlier publications we recommended the termination of the suspicious process; however, from a forensic perspective, this concept may need to be re-thought. For this we need to analyse the requirements for a digital forensic component and compare them to the FATE technology already developed.

3.2 Requirements for an autonomous digital forensics system

Our aim within a FATE with digital forensic capability is to automate the process from the point of incident recognition up to the point where the evidence is securely stored in an archive and can be made accessible to law enforcement entities. For this, we first derive the requirements for such a technology from the current standards, and in particular analyse the RFC3227, which covers best practices and guidelines for evidence collection and archiving. We intentionally do not focus on any particular national or international law to retain the technical character of the paper. In the following subsections we will discuss the key objectives of the RFC3227 sections and identify the requirements we must consider for the implementation of a forensic module in a FATE.

3.2.1 Guiding principles during evidence collection

RFC3227 recommends capturing a picture of the system as accurately as possible, and keeping detailed notes, especially about dates and times. All activities done shall be recorded. It is also recommended that changes to the data during collection be minimised, including access times of files etc. Priority should be given to evidence collection, from the volatile to the non-volatile, and if copies are made, they should be bit copies. The execution of tools in the compromised system should be avoided and distrusted. It must be remembered that interrupting external channels may also result in modifications, including the potential loss of all evidence. Privacy rules should be followed and any intrusion into private data must be limited to the minimum necessary, with reasonable justification. From a legal perspective, all evidence must be admissible, authentic, complete, reliable and believable.

For a FATE system with digital forensic capability, the **accuracy requirement** means a forensic examiner should create a system picture once a system incident is monitored – including date and time information. Since all forensic activities during evidence collection take place on the compromised system, the functionality may not be implemented on the software side, but in the hardware. In that regard, the digital forensic component needs to be a hardware component added to the existing hardware security extensions. This way we can make sure the evidence collection process is assured. By design, the ETPM halts the suspicious process, and modifications by malware inside the process memory space are not possible. Despite this, kernel modules such as the Virtual Memory Manager (VMM) or lower layers in kernel architecture could still access and modify the suspicious process memory space, and therefore it is necessary to run the evidence collection in hardware or to design a trusted kernel driver on the OS side, ensuring that the suspicious memory (and potentially swapped-out parts of a page file) cannot be modified by more highly prioritised system processes (there remains the danger of root kits modification). Instead of a trusted kernel driver, a hypervisor-based module could be considered, especially in a hardware virtual environment (for example, vSentry (Bromium 2015)).

3.2.2 The collection procedure

For the collection procedure, RFC3227 in principle recommends collecting evidence in as much detail as possible. In the incident handling procedures the number of decisions required should also be minimised. The methods used should be transparent and reproducible. In terms of collection steps, RFC3227 sees the need for a plan. For a strong chain of custody, checksums and digital signatures are also recommended.

In terms of collection procedures, the three core requirements of a FATE are the **detail requirement,** the **transparency** requirement, and the requirement for **authenticity**. In an automated autonomous environment such as that in the FATE architecture, it is easy to collect all details of a suspicious process simply by creating a full memory image of it. In addition, the external avenues of the process need to be considered: media access, network connectivity, and inter-process communication. This information can be obtained from the OS or from a trusted kernel driver, as described above in Section 3.2.1. Since the FATE is designed as an open-source hardware environment, we assume transparency is built-in by design. Digital signatures appear to be a good solution to authenticity, and a unique certificate can be securely encapsulated inside the TPM. An ETPM can create its own certificate (derived from its endorsement key), have it signed by a trusted certificate authority (CA), and use it for digital signing.

3.2.3 The archiving procedure

According to RFC3227, evidence must be strictly secured. In addition, the chain of custody needs to be clearly documented, which means it must be clear how the evidence was found and how it was handled – from the time of its collection to the time of its handover to a prosecutor. Access to the evidence must be restricted and must also be clearly documented. Tools should be prepared for storing the evidence and it is preferable that common media should be used (write-once media).

The **chain of custody requirement** appears challenging at first sight, but in fact can be achieved in a FATE without significant effort. It is obvious that archiving information on the compromised system is not an option if modification of the image during its creation cannot be guaranteed (burning it to a DVD could be an example). Due to the link-layer trust relationships between peering systems we can define a protocol extension for the TARP (Maybaum & Toelle, 2015) in order to add an image export functionality to it. Due to its resilience, TARP is a secure method in terms of the chain of custody. A peering system in that context may be configured either as a Forensic Drop Zone (FDZ), which means that the exported images are stored on the peer, or as a Forensic Image Router (FORIR), which is designed to forward the information through the trusted segment to a FDZ peer. The only consideration left is the secure storing of the image in the FDZ. For this, standard procedures of data protection can be applied.

4. Blueprint for a prototype: Trusted forensic module

In this section we introduce a concept for automated autonomous digital forensics in a FATE as a proof of concept, implementing the five identified requirements discussed in Section 3: Trusted Forensics (TFOR). TFOR requires a hardware security extension and an OS security extension that can run autonomously. We demonstrate functionality through the example of an enhancement for the ARM in a proactive ETPM. We will also propose a new kernel layer to accommodate rootkit protection to be used with a passive ETPM.

4.1 Hardware security extension for digital forensics capability

A security incident in a FATE system is first detected by the ISU in the ARM of the ETPM. Before this can happen, however, the ISU must be adapted to forensic requirements. By design, the ISU is supposed to suspend and terminate a suspicious process, but a forensics-capable version needs to keep both the affected process memory space and the external avenues unchanged. To fulfil the accuracy requirement described in Section 3.2.1 it is necessary to halt the process and create the image from the halted unchanged process memory space. In principle, we see two approaches to implementation here:

- Hardware solution

This is the preferable option, since it is the most resilient design, but requires a core redesign of the ETPM: whereas the ETPM is a passive module triggered by the OS through function calls, a proactive ETPM must take action on its own initiative. For this enhancement, the ETPM needs to send high-priority interrupts to stop process execution and start incident handling. For implementation as a TC enhancement, this would require a change in hardware architecture, which needs to be coordinated with the Trusted Computing Group (TCG). New hardware standards would need to be defined;

- Software solution

In a software solution, the current passive ETPM design can be used. We suggest an OS security extension to ensure close to real-time security incident detection by requesting status from the ISU in very short intervals.

Once the suspicious process is halted, evidence collection procedures may start. In our proposal, the additionally required functionality needed for both evidence collection and for the archiving procedure is implemented in a new TFOR module in the ARM (Figure 4); for this, either new intra-module communication channels need to be created or an extended bus message format needs to be designed. An alternative option would be the implementation of TFOR as a new sub-module for the ISU. This option would be efficient since TFOR needs access to the ISU's live status information.

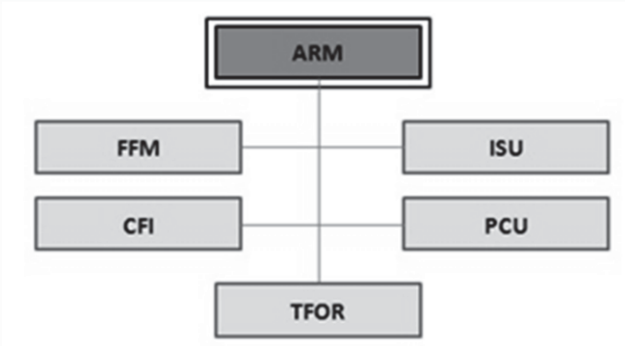


Figure 4: ARM with TFOR module

Such a design would be faster than a separate module which needs to access all information through the ARM-internal bus, but it violates the 'separation of concerns' principle (Dijkstra 1974) since digital forensics is not the core functionality of the ISU. The requirement, discussed in Section 3.2.2, is already fulfilled in the design of the FATE since all details of the security incident are stored in the process memory space, in the page file (swapped memory pages), or in the buffers for the external avenues. Detailed information is thus available. The same applies to the transparency requirement. The open source/open design principle of our solution helps ensure that the entire process of evidence collection is also transparent, simply by design.

The authenticity requirement needs to be discussed in more detail, since this is especially important for legal action to be taken. Every TPM has a unique endorsement key which can be used to generate a pair of keys required for asynchronous encryption methods. From a technical perspective, this would be sufficient proof of authenticity for the image data, but since the endorsement key cannot be exported from the TPM, this missing piece of proof might be a weak point from a prosecutor's perspective. We therefore recommend the use of recognised certificate formats such as X.509 which can be signed by a recognised CA. In a FATE we recommend running a CA at an organisational level or at TIER-1 ISP level. This certificate can be used to digitally sign an image of the suspicious process and would fulfil the authenticity requirement.

For the archiving procedure, we see two viable options for exporting the image to a FDZ, both of which would fulfil the chain of custody requirement:

- Image export based on existing technologies

As long as network communication is still possible, the image can be sent by email or other services using digital signatures to prove authenticity and integrity. Such a message needs to be acknowledged by a receipt which is equally protected and can be validated by the TPM. If no appropriate receipt is received, the network must be considered compromised. From a technical standpoint this requires the implementation of the service used within the TPM – not relying on a potentially compromised OS function;

- Image export by the PCU

Peer communication between the PCUs is considered secure since it is protected by modern asynchronous encryption methods, and due to its implementation at link-layer it is also secure in terms of man-in-the-middle attacks. The image can also be exported using the PCU and avoids the problem referred to in 1. This approach requires a functionality upgrade of the PCU, as well as a protocol extension to the TARP. TARP was designed as a light weight protocol only used for key exchanges and status notifications. To use TARP in the context of image export, a payload option must be added to the protocol. The peering system receiving the TARP traffic must route the traffic forward to the next peer until the FDZ is reached, and serve as a router for images in this context (essentially becoming a FORIR). The FDZ must be configured in an appropriate manner so that only entitled personnel may access the suspicious image. This organisational aspect needs to be further researched, something the authors will elaborate on further in a separate publication.

4.2 OS security extension for digital forensics capability

As long as the ETPM remains a passive component, the automation of the digital forensic process in the TFOR module needs external activation and control. In Maybaum (2015) we described a semi-automated process for control flow integrity monitoring by the ETPM, based on TPM function calls. In particular, it was suggested that the TPM libraries be extended so that with every function call to the TPM an integrity status flag is returned by the ISU and evaluated by the OS. If the integrity status flag indicates a security incident, the OS would terminate the suspicious process and peering systems would be informed about the monitored integrity breach by the ARM (Maybaum & Toelle, 2015).

Although accepted for computing security, this functionality needs to be enhanced for digital forensics. The most important change is giving priority to evidence collection. The suspicious process must not be terminated, but instead be halted and kept in memory. In addition, any external buffers such as media access, IPC, or open network connections, need to be marked, and meta-date kept. A separate kernel module is required for this, which is capable of advising the VMM to lock the affected memory pages, and in a later step to create the image and transfer it to the TPM for signature and export to the FDZ. This is what we call the Trust Management Module (TMM) (Figure 5).

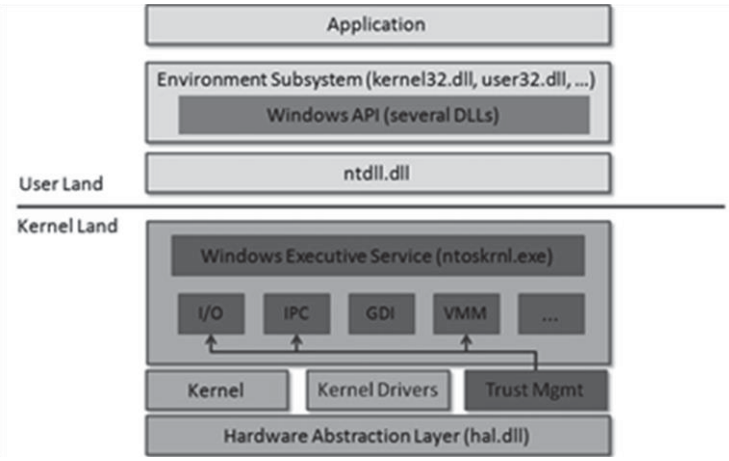


Figure 5: Trust management module in a Windows x32 kernel

This module will also request buffers and meta-data from the external avenues of the I/O and IPC kernel modules. To export the forensic image, the data is forwarded to the ETPM, where it is digitally signed and forwarded to the PCU that will send the image to the FDZ, which will assure the chain of custody. To further enhance the resilience of the FATE we recommend that communication between the TMM and the ARM is also protected by cryptographic means, using (for example) Diffie-Hellman encryption to enable secure communication between the PCUs. TPM function calls can then be enriched by a challenge-response procedure which allows the reliable authentication of the TMM. Since in the system boot sequence the TMM is initiated right after the Hardware Abstraction Layer (HAL), this trust relationship will be difficult to compromise. The only practical option we can imagine for this is a rootkit underlying the HAL layer, where a software solution, as discussed in Section 4.1, may still be vulnerable. This problem does not apply to a hardware solution, however, and such a system could be considered root-kit safe.

5. Conclusions and way ahead

The intention of this paper was to propose a novel idea to further develop the concept of an automated autonomous digital forensic system. We gave an overview of the Future Advanced Trusted Environment concept and introduced two of its core concepts: attack recognition and control flow integrity monitoring and reporting. We discussed our ideas about incident detection and handling in such an environment. To support our suggestions we presented a short discussion of the five main requirements for high quality digital forensics, derived from a best practice standard. In particular we discussed the basic principles of digital forensics with a focus on evidence collection and archiving procedures. Those requirements need to be fully understood for an automated autonomous digital forensic process within a Future Advanced Trusted Environment. Based on these findings we proposed an implementation framework for an automated autonomous evidence collection system. Specifically, we presented a blueprint for a new digital forensics module, 'Trusted Forensics', as a new security

extension for the Attack Recognition Module in an Enhanced Trusted Platform Module. For a robust discussion, hardware and software solutions were evaluated.

We propose using Future Advanced Trusted Environment technology to advance forensic methods of future malware investigation, and persuasive legal evidence based on the Enhanced Trusted Platform Module with its already implemented functionality. The Attack Recognition Module can be extended by the Trusted Forensics module which – as a reaction to a monitored integrity breach – pauses the suspicious process, creates a memory image of the entire process memory space, signs it digitally, and submits it with the help of an extended trusted link-layer network protocol to a trusted peering system, where the image can be stored locally or forwarded to a Forensic Drop Zone. Since this trusted link-layer network protocol can ensure secure communication between Trusted Platform Modules, the memory image is protected against malicious modification, and due to the immediate response to any integrity breach, infiltration of the network connectivity is also unlikely. We see Trusted Forensics as a promising means to securely log any integrity breach identified on a system, and to reliably provide persuasive technical evidence in case of a recognised cyber-attack. We believe that the concept of Trusted Forensics is a significant and necessary step towards automating evidence collection to assist with prosecuting cyber-crime.

We also see remaining risks, however, and a need for further technology enhancements in the future. In a passive design, as seen in the current Trusted Platform Module design, an integrity breach cannot be notified to the operating system in real-time, and the integrity status needs to be requested from the Trusted Platform Module by the operating system. This unfortunately allows a small time frame during which malware could run before the breach is recognised and the execution of the compromised process can be halted. Sophisticated malware could use this time frame to manipulate network connectivity, especially if a targeted attack is made, for example, to disable the export of the suspicious image to an entrusted Forensic Drop Zone. The only option we see to technically ensure the availability of the communication channel to a Forensic Drop Zone is a hardware solution. For this we recommend enhancing the Trusted Platform Module even further by adding a separate isolated communication channel from the module to the network interface card, making it a shared resource between the operating system and the Trusted Platform Module. We have seen this idea successfully implemented within virtual machines. In such a setup, malware that is compromising communication between the system and the network interface card cannot affect the inter-Trusted-Platform-Module-communication and the suspicious image can be directly transferred from module to module without any external interference. Man-in-the-middle attacks can also be identified due to the use of the link-layer for communication, as long as the connected peering system is known.

The biggest challenge to our proposed Future Advanced Trusted Environment technology and the Trusted Forensics capability suggested in this paper is acceptance by the Trusted Computing Group, especially when the technical interface to the system (Low-Pin Count Bus) needs to be changed. This also requires a broad understanding by, and agreement with the hardware industry and operating system manufacturers. Any enhancement of the trusted hardware platform needs to be supported by kernel drivers, as illustrated in Figure 5. As a next step in our research we therefore plan to develop a software demonstrator simulating the Future Advanced Trusted Environment, and capable of emulating security incidents as explained in Section 4. With this simulator the automation of the digital forensic process will be demonstrated practically.

The ideas in Richard and Roussev's article on next-generation digital forensics, which was published 10 years ago, are enhanced by this novel idea. We have seen some of their visions come true, but within the field of automation we have not seen big developments in the last decade. Automation is the next step to advance the field. We see an urgent need to improve the performance of law enforcement in fighting cyber-crime and believe our proposal will contribute to that effort. The Future Advanced Trusted Environment has a great potential. Merging the concept of pro-active Trusted Computing and Digital Forensics can provide the technical means we need to make Trusted Forensics become reality, leading to more prosecutions and, eventually, a safer cyberspace for all.

References

Brenzinski, D. and Killalea, T. (2002) "Best Current Practice Evidence Collection and Archiving", [online], The Internet Society, Network Working Group, Request for Comments 3227, https://www.ietf.org/rfc/rfc3227.txt. [30 January 2016].

Bromium (2015), "Bromium vSentry", [online], http://learn.bromium.com/wp-bromium-vsentry.html.[1].[30 January 2016].
Carrier, B. and Grand, J. (2004) "A hardware-based memory acquisition procedure for digital investigations", *Digital Investigation*, Vol 1, Issue 1, pp 50-60.
Case, A. et al. (2008) "FACE: Automated digital evidence discovery and correlation", *Digital Investigation*, Vol 5, Supplement, pp 65-75.
Casey, E. (2010) *Handbook of Digital Forensics and Investigation*, Elsevier Academic Press, San Diego.
Cohen, F. (2007) *Challenges to Digital Forensic Evidence*, 2nd Edition, ASP Press, New York.
Diffie, W. and Hellman, M. (1976) "New directions in cryptography", *Transactions of the IEEE – Information Theory*, Vol 6, pp 644-654.
Dijkstra E. W. (1982) *On the Role of Scientific Thought, Selected Writings on Computing: A Personal Perspective*, Springer, New York, pp 60-66.
FBI (2016) "Computer Intrusions Combating the Threat", [online], https://www.fbi.gov/about-us/investigate/cyber/computer-intrusions. [30 January 2016].
Maybaum, M. (2015) "Trusted Control Flow Integrity", *Risiken kennen, Herausforderungen annehmen, Lösungen gestalten*, SecuMedia Verlag, Gau-Algesheim, Germany, pp 129-144.[2]
Maybaum, M. and Toelle, J. (2015) "ARMing the Trusted Platform Module", *Military Communications Conference*, MILCOM 2015, IEEE, pp 1584-1589.
Richard, G. and Roussev, V. (2006) "Next generation digital forensics", *Communications of the ACM*, Vol. 49(2), pp 76-80.
Statistica (2015) "Amount of monetary damage caused by reported cyber crime to the IC3 from 2001 to 2014 (in million U.S. dollars)", [online], http://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/. [30 January 2016].
Trusted Computing Group (2016), "About TCG", [online], http://www.trustedcomputinggroup.org/about_tcg. [30 January 2016].
Wang, J. et al. (2011) "Firmware-assisted memory acquisition and analysis tools for digital forensics", *IEEE Sixth International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE)*, 2011, pp 1-5.

[1] White paper – registration required for download
[2] Proceedings of the 14th German IT-Security Congress, May 2015, Bonn, Germany. (In German.)

The Situation Picture in a Hybrid Environment: Case Study of two School Shootings in Finland

Teija Norri-Sederholm[1], Heikki Paakkonen[2] and Aki-Mauri Huhtinen[3]
[1]Department of Health and Social Management, University of Eastern Finland, Kuopio, Finland
[2]Advanced clinical care programme, Arcada University of Applied Sciences, Helsinki, Finland
[3]Department of Leadership and Military Pedagogy, National Defence University, Helsinki, Finland
teija.norri-sederholm@uef.fi
heikki.paakkonen@arcada.fi
aki.huhtinen@mil.fi

Abstract: School shootings are demanding operations for the police, emergency medical services, rescue services, and the Emergency Response Centre. Obtaining a realistic situation picture while the operation is still ongoing, and the perpetrator has yet to be apprehended, is crucial in organising the rescue operation and in minimising further violence. During a shooting incident, the amount of information from different sources can be excessive and there are instances of information overload. The students and staff inside the school and the Emergency Response Centre have valuable information about what is happening inside the building. Police make observations continuously, while searching for the perpetrator(s) and rescuing people. Paramedics and rescue services receive information as to what has happened from those already rescued. It is by no means a straightforward process to collect and collate all the cues and crucial information obtained from the different authorities during the mission, so as to form a common operational picture and to support the development of team situational awareness. However, adequate information management is regarded as a critical success factor in overcoming the challenges of information-sharing during multi-authority missions. Achieving a common operational picture involves taking account of each authority's needs and constraints, thereby enabling the exchange of relevant information between key authorities on-scene and off-scene. According to the literature, sense-making issues between actors with different institutional backgrounds lead to inter-organisational coordination problems. To date, there have been two school shootings in Finland, with a total of 20 fatalities. The incidents occurred in a high school and a polytechnic institution respectively. In this paper, we develop a case study based on official reports and interviews vis-à-vis the two shootings. First, we aim to describe the main elements of information flow during the two incidents in Finland. Next, we analyse the challenges entailed in forming the situation picture on-scene and off-scene in a hybrid environment. Lastly, we discuss the actions implemented to improve information-sharing and the development of the co-operation between the police, the emergency medical services, the rescue services, and the Emergency Response Centre. The latest technology now in use and the role of social media will also be discussed.

Keywords: common operational picture, hybrid environment, information flow, multi-authority co-operation, school shooting, situation picture

1. Introduction

School shootings are demanding operations for authorities such as the police, emergency medical services, rescue services, and the Emergency Response Centre. Obtaining a realistic situation picture while the operation is still ongoing, and the perpetrator is still at large, is crucial to the rescue operation and for minimising further violence. There have been two school shootings in Finland, in Jokela in 2007 (high school) and in Kauhajoki in 2008 (polytechnic), resulting in a death toll of 20. In both incidents, the perpetrators were students of the establishments in question and eventually killed themselves. In Kauhajoki, the perpetrator also started several fires in the building, making the situation even more challenging for the emergency services.

From the operative actors' point of view, a school shooting situation is dangerous and very stressful (Strahler and Ziegert, 2014). Ultimately, the perpetrator usually commits suicide. Perpetrators will typically continue killing victims until they run out of ammunition or cannot locate any more targets (Bradley, 2015). These scenarios are referred to as amok situations, where the violent attack is carried out in a blind rage. It is a question of taking one's own life and the life of others (Strahler and Ziegert, 2014). When the situation is perilous, the role of a situation picture and situational awareness become essential. The situation picture and situational awareness are related concepts with several different definitions. A situation picture can be simply defined as a "subjective snapshot of a certain situation" (Kuusisto, 2005), whereas situational awareness is "knowing what is

going on so you can figure out what to do" (Adam, 1993). Another frequently used term is common operational picture (COP), which is originally a military concept meaning a single display of all collected and combined information shared by more than one command (McNeese et al, 2006). Hence, a COP is a tool for overcoming coordination and information management problems among different organisations and services at different locations during an emergency response (Wolbers and Boersma, 2013), and for enabling situational awareness. All of the aforementioned are derived from information and its interpretation. Therefore, it is crucial to have the right type of information to hand (Endsley, 2000).

During a shooting incident, the amount of information from different sources can be excessive and there are likely to be instances of information overload. The students and the staff inside the school and the Emergency Response Centre have valuable information about what is happening inside the building. The police continuously make observations while searching for the perpetrator(s), and rescuing people. Paramedics and the rescue services receive information as to what has happened from those already rescued. It is by no means a straightforward process to collect and collate all the cues and crucial information received during the mission to form a situation picture, to obtain a COP, and to support the development of team situational awareness. Ojasalo, Turunen and Sihvonen (2009) concluded that in time-critical and rapidly evolving situations, the front line and the first responding law enforcement officers possessed the best situational awareness.

The literature highlights that a recognised challenge concerns the fact that sense-making issues between actors with different institutional backgrounds lead to inter-organisational coordination problems (Wolbers and Boersma, 2013). However, inter-organisational information flow, such as sharing information with other authorities involved in the incident, is the basis for forming the situation picture and when making decisions. If decisions are based on low-grade information, it can lead to poor outcomes and/or risks for the rescue service personnel (McGuinness, 2004). Seppänen et al (2013) collated a list of the major factors that hamper Search and Rescue (SAR) achieving adequate shared situational awareness. Information gaps occur when there is a lack of fluent communication and when COP processes are absent. They also discovered that information gaps occurred when agencies focused only on their own tasks, information delivery processes were unclear, incident information was inadequate, agencies remained passive, and there was a lack of up-to-date information.

Organisational information flow can be described through three phases: sense-making, knowledge creation, and decision-making. First, actors need to understand the changes in the operational environment and their significance. This creates a context for all further actions and guides knowledge creation and decision-making. The knowledge contained in an individual's mind needs to be formulated so that it can be shared and applied. This is about understanding cues and messages. During this phase, actors make choices as to which message to include or exclude, as well as which information to prioritise. The selected information forms possible explanations of what has happened. Actors also share information and opinions, while trying to interpret the situation. Efficient decision-making depends on sense-making. In knowledge creation, the essential questions are: What kind of knowledge is needed and how can it be obtained? During this phase, knowledge is modified by combining both explicit and tacit information. In the decision-making phase, the key action is processing the information and choosing the most appropriate means of achieving the target. The vital question is how to avoid information overload and identify the essential information that needs to be shared with other service participants for decision-making purposes (Choo, 2006).

An important condition for an effective operation involves the face-to-face connection of the on-scene commanders from different authorities and organisations. Situational awareness is improved by establishing an information and communication hub that provides the best possible access to critical information. It also makes the communication swifter and reduces the load on IT and phone networks. Adequate information management has been regarded as a critical success factor in overcoming the information-sharing challenges that emerge during multi-authority missions. Having a COP supports understanding between each of the authorities, identifies their needs and constraints, and enables the exchange of relevant information among key authorities on-scene and off-scene. The role of different technologies, such as information systems, authorities' radio networks, social media, geographical information systems (GIS), radio-frequency identification – RFID-tagged resources, and real-time digital mapping tools, in forming a realistic situation picture is also crucial (Ojasalo, Turunen and Sihvonen, 2003; Stiso et al, 2013.) Many studies relate to different technologies improving the situation picture and situational awareness and generating a real-time COP (Demchak, Griswold and Lenert, 2007; Abdullahi, Qi and Madjid, 2009; Lenert et al, 2011; Artinger et al, 2012; Jokela et al, 2012; Wang, Luangkesorn and Shuman, 2012; and Wu et al, 2013). For example, localisation services provide data on how

many people are inside the building and where they are located. It has been empirically proven that position information is critical in maintaining common situational awareness among a distributed team (Saarinen et al, 2005; Björkbom et al, 2013).

After Jokela, the first school shooting in Finland, the Finnish National Board of Education recommended that every threat should be reported to the police (Investigation Commission of the Kauhajoki School Shooting, 2010). Based on these reports, there have been hundreds of school shooting threats or threatening discourse in Finland since 2007 (Oksanen et al, 2013). As the risk is real, the authorities are now better prepared and have implemented several actions based on experiences in Finland and in other countries. In this paper, we will develop a case study based on official reports and interviews. First, we aim to describe the main elements of information flow during two school shooting incidents in Finland. Next, we analyse the challenges involved in forming the situational picture on-scene and off-scene in a hybrid environment. Finally, we discuss the actions implemented to improve information-sharing and the development of the co-operation between the police, emergency medical services, rescue services, and the Emergency Response Centre. The latest technology currently in use and the role of social media will also be discussed.

2. Material and methods

This paper is based on a qualitative multiple-case study design, which enables differences within the cases to be explored and comparisons to be drawn (Yin, 2009). Empirical data included the official reports by the Investigation Commissions of both school shootings (Jokela and Kauhajoki) in Finland published by the Ministry of Justice, and information gathered from four interviews. The interviewees were chosen from different organisations based on their role and knowledge. The interviews were conducted by the first author and were audio-recorded. The data were subsequently analysed using qualitative content analysis (Krippendorf, 2013). The three themes used in the interviews – information flow, challenges, and actions implemented, including the latest technology – were also used to categorise the data.

3. Findings

3.1 Information flow and situation picture

It was evident from the outset in Kauhajoki that this was a school shooting incident. In this case, the alarm information from emergency calls was delivered to the police, the emergency medical services, and the rescue services. The Emergency Response Centre continuously updated the information received from emergency calls to the authorities, including details to the effect that three people had possibly been shot, the gunman's movements, and the location of possible victims. In Jokela, on the other hand, the initial information was misleading and involved a description of blood coming from a pupil's head, probably as the result of a fall. However, after three minutes, the information was revised and it became evident that a gun had been implicated.

The police patrols inside and outside the building radioed information to other police patrols and to the field commander. They described what they had observed and heard when entering the school building, and when assisting with the evacuation. Although the police patrols inside had the best situation picture, it was not their duty to maintain and update information. The police focused on their main task, namely apprehending the perpetrator, and delivered information at the same time. The primary means of communication was the Public Safety Network (VIRVE). All the authorities used standard internal and co-operation group calls as they would in their normal daily policing operations. This enabled all the authorities to hear what was happening throughout the scenario and helped field commanders keep abreast of the situation picture. The aim was to obtain such a situation picture that the threat could be eliminated as quickly as possible. VIRVE conversations, and the overall operation, were monitored in real time by the National Police Board and the Police University College. However, the fact that VIRVE group calls were used as per daily use disturbed normal communications due to busy radio traffic. Mobile phones were also used and there was face-to-face communication between the authorities. Some information was also obtained from the evacuated students.

In addition to the operative authorities, governmental authorities require a situation picture. They need information in the event that governmental coordination is required for resources. In the Kauhajoki case, for example, the Border Guard Service provided helicopters and the Defence Forces organised helicopters in readiness for transport missions. During incidents of this kind, the public expect the ministries to hold news

conferences to provide official information on the event. Information sent from the scene to the Prime Minister's Office took about 40 minutes to arrive and was delivered through the Ministry of the Interior and the Government Situation Centre. The Government Situation Centre produces a real-time situation picture on the basis of information provided by the competent authorities.

3.2 Challenges in forming a situation picture

Delivering a real-time situation picture in a school shooting case is extremely difficult, and even impossible. Police patrols advancing inside the building and focused on apprehending the perpetrator were working under extreme stress. It was not their responsibility to provide a situation picture. All the police patrols sent information via the authority radio to the field management. The challenge was to pinpoint the relevant information and form the situation picture from the communications. Events unfolded quickly and in the first phase there were insufficient resources to allocate one person to documenting the messages. In both cases, the acute phase was quite short and over before the acute operational command was up and running.

Effectively collecting the information that is essential for rescuing people still trapped inside the school, locating the victims, and finding the perpetrator(s) posed a challenge. There were numerous emergency calls, including new information. Evacuees escaped to nearby buildings until they were eventually instructed to head for the assembly point. They could have provided additional information if they had been interviewed systematically. But in practice this would have been quite difficult as people were in a state of panic, which affected their decision-making capacity. The reliability of their information was questionable, as was the effectiveness of using scarce resources to interview people. However, according to the Investigation Commissions (2008, 2010), no casualties could have been avoided in these cases, even if the authorities had been able to take action sooner. The flow of information among the field commanders represented another operational challenge. A joint field command post for all of the authorities would have been strategic. The various authorities involved did not have a common operational picture and the coordination of their activities was unsystematic to some extent. Knowledge of each other's operational principles and equipment would have been beneficial. Additional people assisting the field commanders would have achieved strength capacity. In the first phase, limited resources and working under high pressure on the front line created additional challenges in delivering the situation picture. One communication challenge that cannot be eliminated in a critical operation is that, for tactical reasons, the police are not able to share all the information.

Technical equipment always constitutes a potential risk in the delivery of information. The fact that activity in VIRVE call groups was exceptionally high during the event caused delays in communications. Further, the authority's network was overloaded. At one point, the access point malfunctioned intermittently and radio units required rebooting, affecting the updating of the situation picture. The mobile phone network was also overloaded at this time. What is more, the sprawling buildings created additional challenges for operations. Making accurate data on the building and the floor plan available to all the authorities would have facilitated a clearer situation picture and aided practical operations. There were also further problems such as the school doors being marked with different symbols from those used in the floor plans. An additional challenge concerns the "front line" people not being able to use computers as communication is carried out via VIRVE.

3.3 Actions implemented to improve information-sharing and co-operation

Several actions have been implemented since these shootings and new tools have been introduced over the ensuing years. The main focus has been on the improvement of information-sharing and co-operation between the authorities in their daily work in order to support the information workflow and situation picture in demanding operations.

The current Emergency Response Centre (ERC) information system has been improved and a totally new system, ERICA, will be implemented during 2016. Nowadays all essential authorities, including hospitals and social care services, are dispatched simultaneously according to a predefined protocol. The dispatch system is able to collect emergency calls related to one incident as a whole. In this way, all the information can be seen in one view, even if several calls are incoming simultaneously. All the information can be shared with the command and situation centres. Notably, in accordance with the law, the content of shared information is based on the authority's role and access rights. However, information overload remains a significant challenge and some information may inadvertently be dismissed. Conversely, if an ERC operator identifies critical information, it can easily be sent to the responsible authority.

The authorities have agreed and implemented new regional and national action models for their daily work in multi-authority incidents and, in principle, the aim is to follow these action models as far as possible. However, there are tailored action models for emergency, hospital, and social care services in the event of a threatened school shooting. There are also detailed instructions for individual actors, depending on their specific role. Added to this, the police have issued special instructions for an active shooter case or a threat, and they conduct regular practice drills based on models tailored to identified threats. The police also have a simplified model, which can easily be communicated to every police officer. The use of VIRVE call groups has been modified and there are now clear principles and instructions on how to use this system in multi-authority incidents. This procedure reduces the load on VIRVE and ensures that important messages are delivered safely. A major lesson learnt in this respect was that when call groups are changed while people are working under extreme stress conditions, there is uncertainty as to whether everyone is actually using the right call group. People had a tendency to continue using their normal daily call groups. In effect, those operating under extreme stress conditions should remain in the call group, while others not experiencing high stress are advised to change the call group.

The Health Care Act (2011) requires that hospital districts are responsible for prehospital emergency care and field commanders are allocated to supervise the area. The readiness for co-operation has been improved significantly due to new organisational and daily action models supporting communication and information flow and improving the situation picture. One vital improvement is the regional 24/7 situation centres for the police, rescue, and emergency services whose responsibility it is to maintain the situation picture. Situation centres have improved the quality of the situation picture and have ensured, for example, that there are dedicated people following VIRVE communications and delivering the information. In a case such as a shooting incident, the situation centre will, in practice, become the operational command centre. When it comes to sprawling buildings, floor plans have now been collected systematically and saved in an information system to achieve a better situation picture.

Information-sharing and co-operation have been improved with the use of new information systems. A field command system for emergency services supporting the regional situation picture has been implemented. In the near future, a common situation picture information system for authorities will be adopted. This will provide a tool for field commanding and for maintaining the common situation picture across the operative authorities. It will reduce the need for radio communication and thus free up the information channels for critical communication. An Unmanned Aerial Vehicle (UAV) is a tool that the authorities have started to use with increasing frequency. With the help of a UAV, it is possible to gain an overview of the area safely and quickly. Mobile technology and the availability of social media have become an innovative way to transmit and receive information. Some rescue departments have successfully used a phone number where people can upload pictures. It is a known fact that even the threat of a school shooting causes activity in social media, on sites such as Facebook and Twitter. In the case of bigger incidents, the authorities in situation centres monitor social media activity. However, it is challenging to detect reliable information when the volume of such information “explodes” through the use of a diverse range of sources. In the case of school shootings, the people involved experience panic while escaping from the perpetrator(s) and it is hard to imagine that they would have the time or inclination to transmit live images using video applications such as Periscope, although outsiders may be in a better position to record the incident. Operative police use body cameras, which are able to send real-time pictures, enabling recipients to gain a more realistic situation picture in the command centre.

4. Discussion

Obtaining a realistic situation picture is crucial for rescuing people, locating victims, and apprehending the perpetrator(s). The two school shootings in Finland provided new insights for the authorities and many improvements have been implemented as a result. However, some factors cannot be changed. For one, school shooting scenarios are invariably dangerous and stressful (Strahler and Ziegert, 2014). The frontline police officers have the best situation picture (Ojasalo, Turunen, and Sihvonen et al 2009), but their task involves advancing inside the building and apprehending the perpetrator(s) under extremely stressful circumstances. As described in this paper, their role does not involve delivering the real-time situation picture. Hence, in the first phase, it has to be accepted that there will be very little information and the situation picture will only become clearer as the operation proceeds.

The fact that there is a common authority network in use in Finland increases the information flow and common operational picture across the authorities. Action models such as the use of VIRVE call groups can prevent delays

and the overloading of communication networks. New regional organisations and the 24/7 situation centres maintaining the situation picture do much to clarify chaotic operations. These centres use video connections for improving communication across different authorities. The use of new technology, information systems, UAVs and body cameras facilitates receiving and processing information to build a more accurate situation picture. However, the use of localisation services inside the buildings remains a challenge and the potential for the role of social media remains questionable in the acute phase of a school shooting operation.

The jointly approved cooperation plans combined with regular training and drilling, as recommended by the School Shooting Investigation Committee, appear to be successful. The system seems to be better equipped both at the operational and governmental level and organisations work more effectively as a result of working action models and tools.

5. Conclusion

School shooting incidents are invariably dangerous, extremely stressful and inherently chaotic. Despite extensive planning and practice, there are a large number of unknown factors. The better the situation picture the operative authorities have, the more effective their ability to control the chaos and the more likely that informed decisions will be made. While frontline police officers have the best situation picture, their task is to apprehend the perpetrator(s) and not to deliver a real-time situation picture. After the two school shootings in 2007 and 2008 in Finland, huge progress has been made in terms of action models, equipment usage, practice drills, training, and in the overall cooperation within organisations and between actors. As a result, a significant step has been taken towards improving the situation picture.

References

Adam, E.C. (1993) “Fighter Cockpits of the Future”, In: Proceedings of 12th IEEE/AIAA Digital Avionics Systems Conference (DASC), October, Texas USA, pp 318–323.

Abdullahi, A., Qi, S. and Madjid, M. (2009) “Situation Awareness in System-of-Systems Ad-Hoc Environments", In: H. Jahankhani, A.G. Hessami and F. Hsu, (Eds.) Global Security, Safety and Sustainability: 5th International Conference, ICGS3 2009, London, UK, September, 2009. Proceedings. New York, New York USA: Springer Berlin Heidelberg, pp 27–34.

Artinger, E., Maier, P., Coskun, T., Nestler, S., Mahler, M., Yildirim-Krannig, Y., Wucholt, F., Echtler, F. and Klinker, G. (2012) “Creating a common operation picture in real time with user-centered interface for mass casualty incidents” [abstract], In: 6th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, Pervasive Health, May 2012, San Diego, California USA. pp 291–296.

Björkbom, M., Timonen, J., Yigitler, H., Kaltiokallio, O., Garcia, J.M.V., Myrsky, M., Saarinen, J., Korkalainen, M., Cuhac, C., Jäntti, R., Virrankoski, R., Vankka, J. and Koivo H.N. (2013) ”Localization Services for Online Common Operational Picture and Situational Awareness”, *IEEE Access,* Vol 1, pp 742–757.

Choo, C.W. (2006) “The Knowing Organization: How organizations use information to construct meaning, create knowledge, and make decisions”, 2nd edition. Oxford University Press, New York.

Demchak, B., Griswold, W.G. and Lenert, L.A. (2007) “Data Quality for Situational Awareness during Mass-Casualty Events”. In: AMIA 2007 Symposium Proceedings November, 2007; Chicago, Illinois USA. pp 176–180.

Endsley, M. (2000) “Theoretical underpinnings of situation awareness: a critical review”, In: M. Endsley and D.J. Garland (Eds.) Situation awareness analysis and measurement. New Jersey: Laurence Erlbaum Associates.

Investigation Commission of the Jokela School Shooting. (2009) “Jokela School shooting on 7 November 2007: Report of the investigation commission, Publication 2009:1, Helsinki: Ministry of Justice, Finland.

Investigation Commission of the Kauhajoki School Shooting. (2010) “Kauhajoki School shooting on 23 September 2008: Report of the investigation commission, Reports and guidelines 39/2010, Helsinki: Ministry of Justice, Finland.

Jokela, J., Rådestad, M., Nilsson, D., Ruter, H., Svensson, L., Gryth, L., Harkke, V., Luoto, M. and Castrén, M. (2012) “Increase situation awareness in major incidents - radio frequency identification (RFID) technique: a promising tool”, *Prehospital and Disaster Medicine,* Vol 27, No. 2, pp 81–87.

Krippendorf, K. (2013) “Content Analysis An Introduction to Its Methodology”, 3rd edition, SAGE, Los Angeles.

Kuusisto, R. (2005) “From Common Operational Picture to Precision Management. Managemental Information Flows in Crisis Management Network”, Publications of the Ministry of Transport and Communications 81/2005, Helsinki, Finland.

Lenert, L.A., Kirsh, D., Griswold, W.G., Buono, C., Lyon, J., Rao, J. and Chan T.C. (2011) “Design and evaluation of a wireless electronic health records system for field care in mass casualty settings”, *Journal of the American Medical Informatics Association*, Vol 18, No 6, pp 842–852.

McGuinness, B. (2004) “Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings”. Command and Control Research and Technology Symposium June, 2004; San Diego, California USA.

McNeese, M.D., Pfaff, M.S., Connors, E.S., Obieta, J.F., Terrell, I.S. and Friedenberg, M.A. (2006) "Multiple vantage points of the common operational picture: Supporting international teamwork", In: *Proceedings 50th Annual Meeting Human Factors and Ergonomics Society*, 2006, pp 467–471.

Oksanen, A., Nurmi, J., Vuori, M. and Räsänen, P. (2013) "Jokela: The social roots of a school shooting tragedy in Finland" In: N. Böckler. Seeger, P. Sitzer, and W. Heitmeyer (Eds.), *School shootings,* pp 189–215, Springer, New York, NY.

Ojasalo, J., Turunen, T. and Sihvonen H.M. (2009) "Responsibility and Decision Making Transfer in Public Safety and Security Emergencies – A Case Study of School Shootings", 2009 IEEE Conference on Technologies for Homeland Security, Boston, USA, May, pp 358–365, Printing House Inc. 10.1109/THS.2009.5168059.

Saarinen, J., Heikkilä, S., Elomaa, M., Suomela, J. and Halme, A. (2005) "Rescue personnel localization system" In: Proceedings 2005 IEEE International Workshop on Safety, Security, and Rescue Robotics, June 2005, pp 218–223.

Seppänen, H., Mäkelä, J., Luokkala, P. and Virrantaus, K. (2013) "Developing shared situational awareness for emergency management", *Safety Science*, Vol 55, pp 1–9. http://dx.doi.org/10.1016/j.ssci.2012.12.009.

Stiso, M.E., Eide, A.W., Nilsson, E.G., Halvorsund, R. and Skjetne, J.H. (2013) "Building a flexible common operational picture to support situation awareness in crisis management", In: T. Comes, F. Fiedrich, S. Fortier, J. Geldermann and T.Müller (Eds.), Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May, pp 220–229.

Strahler, J. and Ziegert, T. (2014) "Psychobiological stress response to a simulated school shooting in police officers", *Psychoneuroendocrinology,* Vol 51, pp 80–91, http://dx.doi.org/10.1016/j.psyneuen.2014.09.016.

Wang, Y., Luangkesorn, L. and Shuman, L. (2012) "Modeling emergency medical response to a mass casualty incident using agent based simulation", *Socioeconomic Planning Sciences*. *Special Issue: Disaster Planning and Logistics: Part 2.* Vol 46, No 4, pp 281–290.

Wayland, B.A. (2015) *Emergency Preparedness for Business Professionals – How to Mitigate and Respond to Attacks Against your Organization,* Elsevier Inc.

Wolbers, J. and Boersma, K. (2013) "The Common Operational Picture as Collective Sensemaking", *Journal of Contingencies and Crisis Management,* Vol 21, No 4, pp 186–199.

Wu, A., Convertino, G., Ganoe, C., Carroll, J.M. and Zhangn, X. (2013) "Supporting collaborative sense-making in emergency management through geo-visualization", *International Journal of Human Computer Studies,* Vol 71, No 1, pp 4–23.

Yin, R.K. (2009) "Case Study Research Design and Methods", 4th edition, SAGE, USA.

The Nexus Between Cyber Security and Energy Security

Daniel Nussbaum[1], Stefan Pickl[2], Arnold Dupuy[3] and Marian Sorin Nistor[4]
[1]Operations Research Department and Energy Academic Group, Naval Postgraduate School, USA
[2] Operations Research, Universität der Bundeswehr München, Federal Republic of Germany
[3]Booz Allen Hamilton contract support to the Deputy Assistant Secretary of Defense for Operational Energy, Washington, USA
[4]Department of Computer Science, Universität der Bundeswehr München, Federal Republic of Germany
dnussbaum@nps.edu
stefan.pickl@unibw.de
arnold.c.dupuy.ctr@mail.mil
sorin.nistor@unibw.de

Abstract: The centrality of cyber and the issues surrounding it are incontrovertible in today's world. This is true across practically every domain of human behavior, of which some examples are warfare, business, and economics. What is less recognized are the critical linkages that exist between the incontrovertible centrality of cyber and other critical resource issues, three examples of which are water supply, transport security and especially energy security. We propose a special graph-based approach which can be used for all of these critical infrastructures.

Keywords: energy security, cyber security, energy-cyber nexus, critical infrastructure, graph-based approach

1. Introduction

This paper addresses in the context of critical infrastructure protection one of these critical cyber connections, namely the issue of the nexus between Cyber Security and Energy Security. In particular, the paper addresses the following subjects:

- Characterization of Energy
- Transport Security in the context of Cyber Security

In this specific triangle we analyze the following intersection:

- Which specific problems exist in the intersection between Energy Security and Cyber Security?
- What research roadmaps have already been published to address these problems that lie in the intersection?
- Which portions of these roadmaps, if any, have already been addressed or accomplished?

It is hoped the best that this paper serves as an opening for a rich and continuing conversation on the critical links between these three important issues, Cyber Security, Transport and Energy Security.

2. Energy security

Energy Security is a phrase that is used by many people and with many meanings, but it is worth noting that there is no standard, universally accepted, definition of this phrase. Of course, within the national Security arena, we care because of the impacts on lives, mission, and cost.

2.1 Examples concerning energy security

In the following, we present some motivating examples:

- 3,000 US service members have been killed or injured defending fuel and water supply lines in Iraq and Afghanistan. (2011)
- Marines increased time in the field from 3 days to 3 weeks and reduced weight by 30 lbs./13 kg when they fueled with solar panels instead of batteries.

- Every $1.00 increase in price of a barrel of petroleum costs the Navy $31M in unbudgeted funding annually, coming out of readiness accounts (e.g. training, flight hours)

Figure 1: Taliban bomb attack on tankers carrying fuel to NATO forces. Copyrights of (The Telegraph and Farmer, 2012)

Figure 2: Fuel convoy, Afghanistan. Copyrights of (Deloitte, 2010)

2.2 Critical infrastructure protection: Energy sectors

According to literature the following 8 paragraphs, through "Prof G. Bahgat, of the US National Defense University…", are reproduced from (Nussbaum, March-2014).

All sectors of society and all sectors of the economy rely on energy, so that disruptions to energy have important consequences across the full spectrum of organizations, populations, and systems. *Energy Security is characterized as the set of conditions in which energy is able to fulfill the demands placed on it by the society and economy.*
It is also the case that Energy Security is essential to national defense, including the conduct of military operations. Yet achieving Energy Security is challenging because of its complexity and extent.

Activities as diverse as the burning of Iraqi oil fields after the first Gulf War, natural disasters which lead to catastrophes such as Fukushima or Hurricane Katrina, and attacks on electrical substations all have far-reaching impacts. Moreover, these diverse challenges to Energy Security affect different portions of society and the economy, and they therefore require different preventative and ameliorative mechanisms.

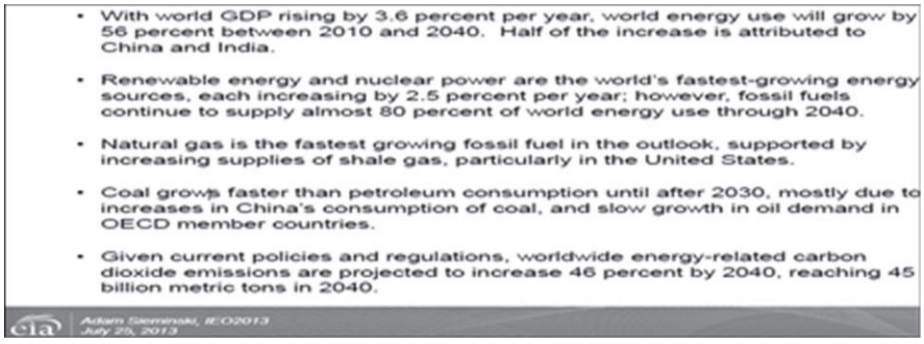


Figure 3: The International Energy Office's key findings of the International Energy Outlook 2013

To provide some background and context, it is important to note that there is a greatly expanding use of energy in the world, and therefore the opportunities to disturb Energy Security are greater now than they have been in the past, and they are less now than they will be in the future. ***Error! Reference source not found.*** and *Figure 4* display, respectively, the International Energy Office's Key findings of the International Energy Outlook 2013,

and the estimated growth in worldwide energy demand over the next 20+ years. Both of these figures provide clarity to the assertion that growth in total energy demand, including in the components of that demand, are poised for significant growth, as shown in the graphic below.

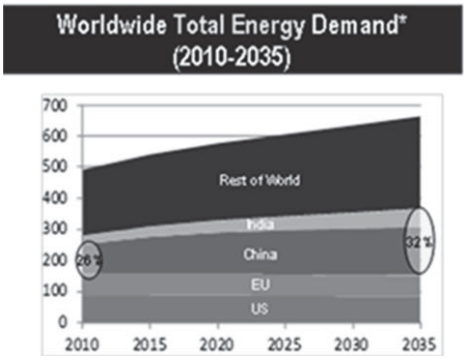


Figure 4: The estimated growth in worldwide energy demand over the next 20+ years

In (Elizabeth Rosenberg, February-2014) Energy Security for the United States is defined to mean "…*reliable access to sufficient, affordable energy supplies to fuel economic growth".*

2.3 Energy security and the "4 A's"

Another well-known description of Energy Security, known as the "4 A's", defines Energy Security in four dimensions:

- Availability. This dimension speaks to the physical existence of energy. It may be either easy to retrieve, or, as in the case of "unconventional oil", it may be hard to retrieve.
- Accessibility. This dimension speaks to the geopolitical environment and, in particular, as defined in the paragraph above, addresses whether the energy that is "available" can be retrieved and made usable.
- Affordability. This is the financial dimension of the Energy Security issue, and it has the usual connotations of whether the ability to obtain and use the energy is within one's budgetary means. However, upon further inspection, this concept can be complex and not as easily understood as intuition would suggest. Indeed, it is not well-defined in the literature, although it has an intuitive meaning. The interested reader can see some important work in this area by looking through either of the websites identified below or in the ***Error! Reference source not found.***. Especially in the document entitled "Affordability Research Document" (MORS, 2014), which asserts that "*Affordability is an abstract term that most people think they understand but have difficulty defining or explaining."*
- Acceptability. This addresses the issue of environmental acceptability and stewardship, in the sense that the environmental footprint associated with Accessing the energy does not violate some set of policy – based norms.

Figure 5: MORS held a workshop on affordability analysis: How do we do it? on 1-4 October 2012

3. Cyber security in the context of energy security

In (Bahgat, 2011) Cyber Security is a phrase that is used by many people and with many meanings, but it is worth noting that there is no standard, universally accepted, definition of this phrase. It covers all aspects of ensuring the protection of IT critical infrastructures, and it further includes the notion of Resilience, which is the ability to protect and repel and recover from cyberattacks. Common phrases like (Wikipedia, 2016) define "*Computer security, also known as cybersecurity or IT security, is the protection of information systems from theft or damage to the hardware, the software, and to the information on them, as well as from disruption or misdirection of the services they provide. It includes controlling physical access to the hardware, as well as protecting against harm that may come via network access, data and code injection, and due to malpractice by operators, whether intentional, accidental, or due to them being tricked into deviating from secure procedures.*"

Cyber security standards are therefore security standards which enable organizations to practice safe security techniques to minimize the number of successful cyber security impacts. Cyber security refers to the technologies and process each designed to protect computers, networks and data from unauthorized access, vulnerabilities and attacks delivered via the Internet. A well-known aphorism in this arena is "*The only system which is truly secure is one which is switched off an unplugged"*

Referring to (Scott Jasper, March-2014) cyber warfare is a "conflict between states where precise and proportionate force is directed against military and industrial targets for the purposes of political, economic or territorial gain." "Cybered conflict" could be considered a better term than 'cyberwar" to frame the complexity and ambiguity of struggle involving cyberspace, including asymmetric conflicts, hybrid warfare and counterterrorism campaigns.

3.1.1 Cyber security and the comprehensive approach

The Office of Cyber Policy in the US Secretary of Defense confirms the need for the US Federal Bureau of Investigations, the Department of Homeland Security and the Department of Defense to coordinate with public, private and international partners in cyber security efforts. Although some workshops have occurred on a comprehensive approach to cyber security and others on cyber deterrence, there is little empirical work on the intersecting issues.

The questions and points in this paper offer a start point to analyze the viability of offensive concepts, defensive measures and cooperative mechanisms. Further research is necessary to determine exactly how a comprehensive approach can achieve complementary strategies for deterrence of cyber aggression. Here, we propose a specific graph-theoretic approach which can also be used for energy and transportation networks.

3.2 Sector models within the comprehensive approach

All sectors of the economy rely on the networks, systems, and services that form the integrated and interconnected domain known as cyberspace. Information and Communication Technologies are essential to national defense including the conduct of military operations. Yet protecting the cyber domain is challenging because it is boundless, subject to change, and open to all comers. Recent events reveal that cyber aggression is relentless, pervasive, and dangerous. Acts of aggression include theft or exploitation of data; disruption or denial of access or service; and destructive action including corruption, manipulation, and damage.

Cyberspace is probed and penetrated by hackers, criminals, terrorists and foreign powers. As an interdependent network of Information and Communication Technology infrastructures, it is insufficient to differentiate between commercial, civil and military spheres in cyberspace. For example, private industry owns and operates 90% of the critical infrastructure in the United States. Cyberspace integration brings new levels of vulnerability and potential for cascading disruption of critical infrastructures or key resource sector functions. While cyberspace relies on the servers, switches, and routers of individual countries, this digital structure is globally connected and global cooperation is required to secure it.

Cyber warfare is defined in US joint terminology as "an armed conflict conducted in whole or part by cyber means." An attacker could launch a military confrontation during a period of tension by attacking civilian infrastructure, a cyber-attack just prior to or simultaneously with a surprise military attack, or wait until war starts to activate implanted exploits. Acts of cyber aggression, such as criminal exploitation, military or industrial

espionage, nationalist hacker protests, and infrastructure infiltration or sabotage seen today, might represent lower level means of cyber warfare. The buying or renting of viruses (malicious code), exploits (of code vulnerabilities), bot (compromised machine) networks, and command and control servers provide an array of tools and services for motivated threat actors and states. In addition to "military operations to deny an opposing force the effective use of cyberspace systems and weapons," some nation state cyber campaign doctrine appears to include disruption of governmental services, financial enterprises, and media outlets.

The Distributed Denial of Services (DDoS) assaults (that flood systems with useless traffic) upon Georgian infrastructure in 2008 heralded the reality that cyber aggression will be a component of any future conflict. Russian blogs and forums spread instructions and script to patriotic hackers to disrupt Georgian public and private sector Web sites.

3.3 Command and control: Logistics networks and power grids

During the ground invasion, command and control servers managed by a cybercrime group issued DDOS attack commands. Similar tools and tactics were used by Russian nationalists in the cyber riot in Estonia in 2007. Other nations might consider comparable uses of non-military actors and services against civilian targets in obtaining their objectives.

Military doctrine in China calls for attacks on the critical infrastructure of an opponent's homeland in case of conflict. For a conflict over Taiwan, Chinese computer network operations strategy appears designed to target US logistics chains in host nations in the region, as well as continental US logistics networks and companies. Computer-induced failures of power grids, transportation networks, or financial systems could cause physical damage and economic disruption. US military operations, both at home and abroad, are dependent on this critical infrastructure. Chinese capabilities could impede military readiness and the operation of US critical infrastructure. Documented cyber aggression demonstrates we may already be in phase zero (the beginning) of cyber warfare. The Pentagon made such allegations against China in its annual report, alluding to the use of "computer network exploitation capability to support intelligence collection against the US diplomatic, economic, and defense industrial base sectors that support US national defense programs. Exposure by an American company of a hacking campaign based out of Shanghai focused on drone technology is the latest confirmation of the Washington Post published list of military systems compromised by cyber espionage emanating from China.

4. Modelling and analysis of the vulnerability of cyber attacks

Energy production and distribution resources, as well as energy facilities are vulnerable to cyber-attack. We might well ask "why weren't these energy resources designed to be inherently resistant to cyber-attacks?" The over-riding and obvious reason that these energy resources were not designed to be inherently resistant to cyber-attacks is that many of these assets were developed before cyber threats were an issue, and it was only later, after these energy resources had been designed, built, and put into use, that these assets were networked together. Many stakeholders, within the government, as well as in the private sector and in academia, have taken notice of the vulnerability to cyber-attacks, and the consequences of such attacks, on energy networks.

While there are efforts to standardize definitions, regulations, topologies, and operational guidance for cyber security, and to promote the adoption of new practices and technologies, these efforts tend to be voluntary. Additionally, private energy producers are incentivized to be proactive when addressing cyber threats, so that they can identify, address, and obviate potential issues and problems, whether they are brought to their attention from their stakeholders, including shareholders and government agencies.

4.1 SCADA supervisory control and data acquisition versus DCS distributed control systems

The two main types of energy delivery systems are the Supervisory Control and Data Acquisition (SCADA) and Distributed Control Systems (DCS). Even though the Integration of these SCADA systems into networks in order to provide for better monitoring and distribution has increased their vulnerability to cyber threats, early SCADA system designs did not account for potential future cyber threats.

In 2014, U.S. News reported "*only 32 percent of electric utilities surveyed for the report had integrated security systems with the 'proper segmentation, monitoring and redundancies' needed for cyber threat protection.*

Another 48 percent said they did not" (U.S. News and Neuhauser, 2014). (The Wall Street Journal and Smith, 2014) mentions that most gas and electric utilities were still running on the vulnerable Windows XP operating system, and that the entire nation could experience a blackout due to successful attacks on just 9 out of 55,000 electrical substations.

Given the issues, as stated in the preceding paragraphs, we ask "who, if anyone has addressed these issues?"

4.2 Cybersecurity for energy delivery systems (CEDS)

The following instances summarizes the statements of the last paragraph:

- Energy Reliability. The US Department of Energy's (DOE) Office of Electricity Delivery and Energy Reliability asserts, on its website (U.S. Department of Energy, 2011), that it had "*designed the Cybersecurity for*

Energy Delivery Systems (CEDS) program to assist the energy sector asset owners (electric, oil, and gas) by developing cybersecurity solutions for energy delivery systems through integrated planning and a focused research and development effort. CEDS co-funds projects with industry partners to make advances in cybersecurity capabilities for energy delivery systems."

- Cyber Emergency Response Team. Similarly, the US Department of Homeland Security's asserts that its Industrial Control Systems Cyber Emergency Response Team (ICS-CERT, 2016) "*works with government and private industry to reduce risks to critical infrastructure*" (Perez et al., 2015).

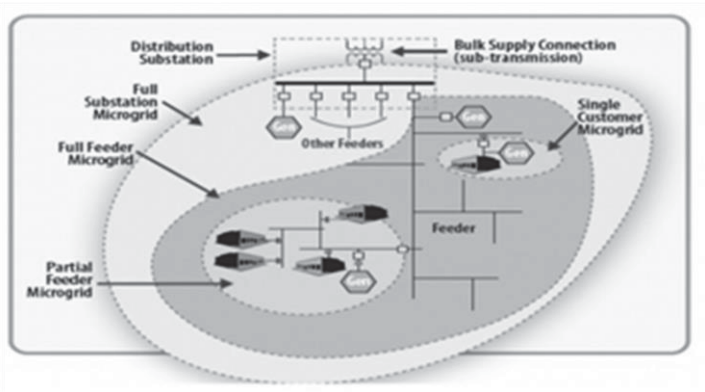


Figure 6: The role of microgrids in helping to advance the nation's energy system (U.S. Department of Energy, 2012)

- RCES: Regional Cyber & Energy Security Center. The University of Texas at El Paso asserts that its Regional Cyber & Energy Security (RCES, 2013) Center at "*develops cyber and Energy Security capabilities addressing the technical, regulatory, and commercialization challenges. The RCES Center also stimulates opportunities associated with the intersection of cyber, cyber-physical, and Energy Security technologies, particularly the increasing overlap of information technology (IT) and operational technology (OT*)". It is the authors' view that RCES is a rich and textured resource for understanding the intersection of the Energy Security and Cyber Security domains.

4.3 Proposed solutions in the intersection: Risk cycle

The US DOE Office asserts on its website that its Cybersecurity for Energy Delivery Systems (CEDS) program is working with stakeholders in the energy sector to develop cybersecurity standards and practices. Specifically, "*CEDS program activities fall under five project areas, guided by the Roadmap to Achieve Energy Delivery Systems Cybersecurity.*

These five areas which can be seen as a certain risk cycle are:

- Build a Culture of Security. Through extensive training, education, and communication, cybersecurity "best practices" are encouraged to be reflexive and expected among all stakeholders.
- Assess and Monitor Risk. Develop tools to assist stakeholders in assessing their security posture to enable them to accelerate their ability to mitigate potential risks.

- Develop and Implement New Protective Measures to Reduce Risk. Through rigorous research, development, and testing, system vulnerabilities are revealed and mitigation options are identified which has led to hardened control systems.
- Manage Incidents. Facilitate tools for stakeholders to improve cyber intrusion detection, remediation, recovery, and restoration capabilities.
- Sustain Security Improvements. Through active partnerships, stakeholders are engaged and collaborative efforts and critical security information sharing is occurring."(U.S. Department of Energy, 2011).

Additionally, there are other schools and organizations that are coordinating the development of improved software and practices to be implemented by the energy sector.

5. Cybersecurity risk management programs

From the Data Protection Report (Perez et al., 2015): "A significant step in the standardization of cybersecurity protocols for the energy sector came in January 2015, when the US Department of Energy, Office of Electricity Delivery and Energy Reliability issued the Energy Sector Cybersecurity Framework Implementation Guidance. This guidance is intended "to help the energy sector establish or align existing cybersecurity risk management programs to meet the objectives" of the US National Institute of Science and Technology (NIST) Cybersecurity Framework. The Cybersecurity Framework itself is currently voluntary, but it is already being implemented by businesses. It bears mentioning that the energy sector has experienced an evolution of the North American Electric Reliability Corporation's (NERC) Reliability Standards, from a voluntary standard before the 2003 blackout, to a mandatory requirement after the event."

Indeed, for many companies, the potential for public censure alone provides enough of an incentive to take prudent compliance steps. For public companies, the additional exposure to potential shareholder suits effectively makes heeding the government's call to action almost compulsory. The highest priority for energy companies wishing to mitigate their potential liability for cybersecurity breaches is understanding which regulators might come looking for answers in the event of a cyber incident, and what information and documentation they will seek. The second step is understanding government expectations with regard to prioritization of energy assets and their associated risk. The second step is necessary for cost-effective compliance program development and implementation. Notably, owners and operators of energy infrastructure assets are at higher risk of being targeted for investigations because, often, they are the first and last line of defense against cyber-attacks.

6. Conclusion: DoD perspectives of the cyber-energy nexus – leadership programs

6.1 Platform information technology PIT

The experience of the US Department of Defense (DoD) in the cyber-energy nexus is perhaps not too different from other Government organizations or the private sector. However, the scale of the problem is significantly larger and growing with clear implications to national security. Broadly speaking, DoD spends hundreds of millions of US dollars on cyber security for information systems (IS) and to mitigate the national security threats from millions of associated IS devices. Existing in the IS sphere is Platform Information Technology (PIT), defined by DODI 8500.01 as "*…both hardware and software, that is physically part of, dedicated to, or essential in real time to the mission performance of special purpose systems*" (NAVAIR, 2007).

6.2 Industrial control systems ICS

It is estimated there are tens of millions of PIT systems and their associated devices currently in use within DoD alone. Private sector PIT vulnerabilities have been exploited for years, and attackers are increasingly sophisticated at compromising PIT systems and their associated devices. In fact, DoD has yet to measure the scope and magnitude of the threats to which it may be exposed through its PIT environment. Delving deeper into the PIT realm, Industrial Control Systems (ICS) are a subcomponent which encompass multiple control systems, with direct applicability to the security of the operational energy infrastructure. Within DoD, PIT includes ICS, so for the purpose of this analysis, the terms are used interchangeably. ICS include supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and other systems such as Programmable Logic Controllers (PLCs), which are frequently used in industry and critical infrastructures. For DoD, ICS is employed in its broadest sense and represents the full range of control systems (SCADA, DCS,

building, vehicle, transportation, etc.) located in DoD facilities. Moreover, ICS are often installed piecemeal using commercial off the shelf (COTS) components, frequently installed by varying contractors using non-standard equipment. ICS may be a loosely connected system of systems, typically consisting of a multi-facility front end, an installation-wide IP network, and multiple subsystems, all of which contributes to the broader environmental vulnerabilities to outside penetration.

6.3 Cyber command

DoD is addressing all five elements of the DOE roadmap, notably through Cyber Command and the efforts of the individual Services. However, from a strictly operational energy context, it takes a narrower viewpoint primarily from the prevention, as opposed to the mitigation or recovery standpoint.

7. Outlook: DoD leadership and complex decision making processes

There is now an effort underway to address these challenges by focusing on the operational energy infrastructure critical to DoD missions. The overarching objective of this research effort is to quantify the problem of cyber security for energy-related platform information technology (PIT)/industrial control systems (ICS) to aid DoD leadership in their decision making process. This research will provide senior leaders with a focused analysis of PIT exploitability with which to enhance and enable appropriate resource decisions.

The above-mentioned analysis will focus on the operational energy infrastructure critical to DoD missions. More specifically, there will be focus on three key areas: 1) identifying exiting vulnerability to known threats, 2) assessing technological capabilities and identifying gaps, and 3) identifying workforce skillsets to meet the operational energy challenges going forward. Additionally, the effort will baseline the typical DoD configurations, protocols, and threat exposures of fixed and tactical infrastructures, notably, utilities, energy facility systems, microgrids, prime/standby power generation, vehicle recharging stations, physical security systems, fueling systems and reverse osmosis units.

7.1 Management cockpit for energy-transport-cyber network analysis

Analyzing the vulnerability of a network and identifying critical spots is of great importance for today's decision makers. The in-depth knowledge about the underlying network and its efficiency is fundamental to adequate decision making. So far, the usability of networks was analyzed via individual measures that considered shortest routes or bottlenecks. As an extension to this in (Nistor et al., forthcoming), the flow-weighted efficiency measure was introduced and exemplary demonstrated on a physical transportation network – the underground network of Munich, Germany.

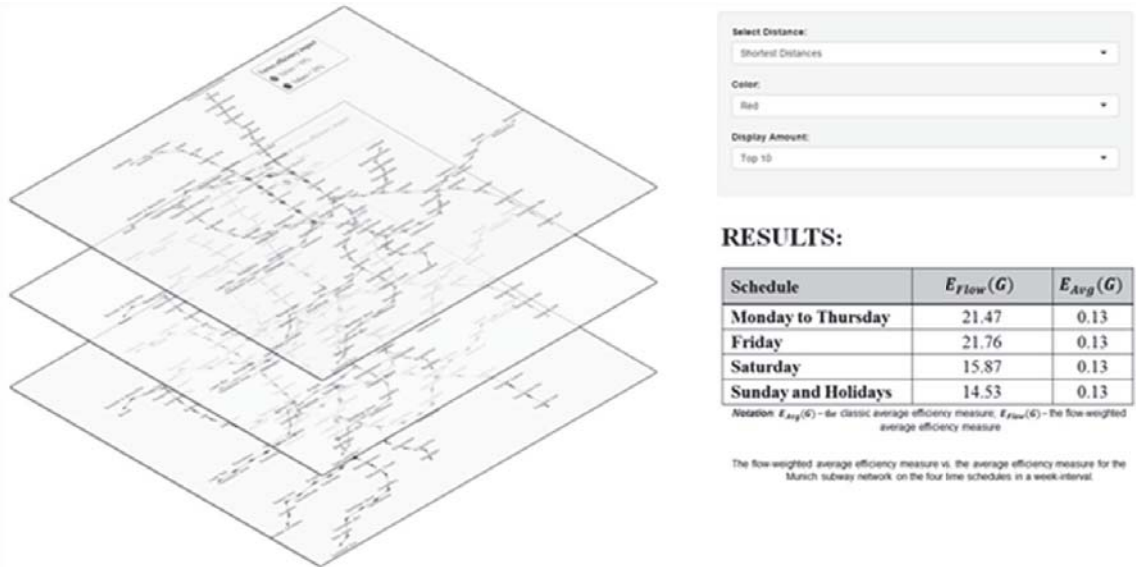


Figure 7: Example of a multi-layer management cockpit for the energy-transport-cyber network analysis of the underground network from Munich, Germany

This paper addresses the general usability of the weight values of a graph from an efficiency point of view: For transportation as well as energy networks and especially also for cybersecurity concerns. The proposed measure

in (Nistor et al., forthcoming) calculates the flow-weighted efficiency in a subway network by computing the shortest route between every pair of stations and the according bottleneck flow of the trains. Results show that the network efficiency is invariant over all schedules, whereas the flow-weighted efficiency is significantly varying; it is highest on the densest schedule and lowest when the least trains are operated. This invariance is a great advantage for general analytic tasks. *Figure 7* displays an example of how a multi-layer energy-transport-cyber network could be analyzed in the management cockpit of a physical transportation system like the underground network of Munich, Germany.

7.2 Security management and architectures

By means of this proposed graph-based analysis it is possible to optimize the quality management of such systems which were discussed in this paper in terms of maximizing serviceability via various security installations and back-up solutions. We embed these analytic tools in a general Management Cockpit DSS.

7.3 Management cockpits for critical infrastructure protections MACCI

The military benefits of this management cockpit vision will address challenges and provide leaders with a snapshot to help in the decision making process. It will provide a repository of best practices, security architectures, security controls and/or compensation controls that increase resilience to known attack tools. Moreover, this specific analysis will provide solutions for both operational energy and installation energy missions, and help guide future energy-related research and development.

Finally, it will improve the collaboration and professionalism of DoD's PIT/ICS stakeholders and workforce.

Acknowledgements

Research of author Marian Sorin Nistor, was funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA Grant Agreement Number 317382.

References

Bahgat G (2011) *Energy security: an interdisciplinary approach*: John Wiley & Sons.

Elizabeth Rosenberg (February-2014) *Energy Rush: Shale Production and U.S. National Security*. Report of the Unconventional Energy and U.S. National Security Task Force.

ICS-CERT (2016) *Industrial Control Systems Cyber Emergency Response Team*. Available at: https://ics-cert.us-cert.gov/ (accessed 20 May 2016).

MORS (2014) *Affordability Research Document*. Available at: http://www.mors.org/UserFiles/file/2013-Affordability-Analysis/Affordabilty_Analysis_Research%20_%20v%2023_2014-01-27.pdf (accessed January, 2014).

NAVAIR (2007) *Platform information technology definitions for the department of the navy*. Available at: http://www.navair.navy.mil/nawctsd/Resources/Library/IA/Files/Enclosure_1_-_Platform_IT_Definitions_for_DON.doc (accessed 20 May 2016).

Nistor MS, Pickl SW, Raap M and Zsifkovits M (forthcoming) Network Efficiency and Vulnerability Analysis using the Flow-Weighted Efficiency Measure. In: *EMNet-BOOK: Management and Governance of Networks*.

Nussbaum DA (March-2014) A Comprehensive Approach to the Challenges of Energy Security: An Overview with an example drawn from the Continuum of Military Operations.

Perez T, Segalis B and Navetta D (2015) *Energy cybersecurity – a critical concern for the nation*. Available at: http://www.dataprotectionreport.com/2015/04/energy-cybersecurity-a-critical-concern-for-the-nation/ (accessed 20 May 2016).

RCES (2013) *Regional Cyber and Energy Security*. Available at: http://rces.utep.edu/ (accessed 20 May 2016).

Scott Jasper (March-2014) A Comprehensive Approach for Deterrence of Cyber Aggression.

The Wall Street Journal and Smith R (2014) *U.S. Risks National Blackout From Small-Scale Attack*. Federal Analysis Says Sabotage of Nine Key Substations Is Sufficient for Broad Outage. Available at: http://www.wsj.com/news/articles/SB10001424052702304020104579433670284061220 (accessed 20 May 2016).

U.S. Department of Energy (2011) *Energy Delivery Systems Cybersecurity*. Available at: http://energy.gov/oe/services/technology-development/energy-delivery-systems-cybersecurity (accessed 20 May 2016).

U.S. News and Neuhauser A (2014) *Cybersecurity Among Top Energy Industry Concerns* (accessed 20 May 2015).

Wikipedia (2016) *Computer security*. Available at: https://en.wikipedia.org/wiki/Computer_security (accessed 20 May 2016).

The Automated Detection of Trolling Bots and Cyborgs and the Analysis of Their Impact in the Social Media

Jarkko Paavola[1], Tuomo Helo[1], Harri Jalonen[1] Miika Sartonen[2] and Aki-Mauri Huhtinen[2]
[1]Turku University of Applied Sciences, Turku, Finland
[2]Finnish National Defence University, Helsinki, Finland
jarkko.paavola@turkuamk.fi
tuomo.helo@turkuamk.fi
harri.jalonen@turkuamk.fi
miika.sartonen@mil.fi
aki.huhtinen@mil.fi

Abstract: Social media has become a place for discussion and debate on controversial topics, and thus provides an opportunity to influence public opinion. This possibility has given rise to a specific behavior known as trolling, which can be found in almost every discussion that includes emotionally appealing topics. A troll is an individual who shares inflammatory, extraneous or off-topic messages in social media, with the primary intent of provoking readers into an emotional response or otherwise disrupting on-topic discussion. Trolling is thus a useful tool for any organization willing to force a discussion off-track in the situations when one has no proper facts to back one's arguments. In this paper, the analysis of trolling is based on public discussion stakeholder classification by Luoma-Aho (2015), including positively engaged faith-holders, negatively engaged hateholders, and fakeholders. Trolls can be considered as either hateholders (humans) or fakeholders (bots or cyborgs). It is stated by Luoma-Aho that the influence of a fakeholder appears larger than it really is in practice, but tools for analyzing the impact are not provided in her work. This paper continues the work by Paavola and Jalonen (2015), who examined in their paper whether sentiment analysis could be utilized in detecting trolling behavior. It was concluded that sentiment analysis as such cannot detect trolls, but results indicated that social media analytics tools can generally be utilized for this task. In this paper the work continues with automatic detection of bots, which facilitates the analysis of fakeholder communication's impact. The automatic bot detection feature is implemented in the sentiment analysis tool in order to remove the noise in a discussion.

Keywords: social media, stakeholder, trolling, sentiment analysis, bot, cyborg

1. Introduction

The rise in diverse Internet threats has opened up the discussion on the possibility for nation states to extend their capacity to control information networks, including citizens' private communications. Due to the plethora of information available, people are not always able to recognize whether information is valid or not, and consequently tend to make hasty presumptions with the data they have. This tendency is utilized by "trolling", which the media has equated with online harassment in recent years. Because of trolling, it is becoming increasingly difficult to pinpoint where information originates from and where it leads to (see Malgin 2015).

We are at a stage in the evolution of information in which the unpredictability of its effects is accelerating. The volume of information is growing, and its structure is becoming increasingly opaque. Information can no longer be seen as a system, or as the extent of one's knowledge, but more as an entity that has started to live a life of its own. Information both provides its own energy and is its own enemy. Certainly, information is also a source of good development and can improve our quality of life. It is essential, however, to understand that it can also unleash danger and adversity.

Information is essentially a product of engineering science. In order to expand our sphere of understanding with information as a part of human social life, one has to step outside the "hard sciences" realm. Social scientific point of view is especially called for when discussing possible threats and human fears connected with information.

Computer culture theorists have identified the richly interconnected, heterogeneous and somewhat anarchic aspect of the Internet as characterizing a social condition that is rhizomic in nature (Coyne 2014). During the past quarter of a century the usefulness of the Internet has permeated all domains (individual, social, political, military and business). Worldwide, everyone can use the Internet without any specific education. We do not have to concern ourselves with the technical aspects of the Internet, or devise detailed plans on how to communicate in social media. At the same time, our work-related and official messages run parallel with our

private communications. Similarly, our emotions and rational thinking may easily become intertwined due to the ease and immediacy of our communications. Deleuze and Guattari (1983) use the terms "rhizome" and "rhizomatic" to describe this non-hierarchical, nomadic and easy environment, particularly in relation with how we behave in this kind of environment.

The intercontinental network of communication is not an organized structure; it has no central head or decision-maker; it has no central command or hierarchies. The rhizomatic network is simply too big and diffuse to be managed by a central command. By the same token, rhizomatic organizations are often highly creative and innovative. The rhizome presents history and culture as a map, or a wide array of attractions and influences with no specific origin or genesis, for a "rhizome has no beginning or end; it is always 'becoming' in the middle and between things" (Deleuze & Guattari 1983). One example of the diversity of rhizome networks is the fakeholder behavior. In order to have a more thorough view of the discussion, it would be important to know the sources behind the fakeholders' arguments, but like the artists of black propaganda, they attempt to hide themselves. It can be hypothesized that the role of fakeholders increases with subjects whose legitimacy is questioned or challenged, and when the public is confused about the relevance and significance of the arguments presented in various social media platforms.

The promise of social media is not confined to technology, but involves cultural, societal and economic consequences. Social media refers herein to a constellation of Internet-based applications that derive their value from the participation of users creating original content, modifying existing material, contributing to a community dialogue and integrating various media together to create something unique (Kaplan & Haenlein 2010). Social media has engendered three changes: 1) the locus of activity shifts from the desktop to the web, 2) the locus of power shifts from the organization to the collective and 3) the locus of value creation shifts from the organization to the consumer (Berthon et al. 2012)

Social media has been integrated with the lives of postmodern people. Globally, over two billion people use social media on a daily basis. Whether it's a question of the comments of statesmen, opposition leaders' criticism or celebrities' publicity tricks, social media offers an authentic information source and effective communication channel. Social media enables interaction between friends and strangers. It lowers the threshold of contact and personalizes communication. In a way, social media has made the world smaller.

Social media has brought with it 'media life', which Deuze (2011) calls "the state where media has become so inseparable from us that we do not live with media, but in it" (Karppi 2014, 22). In a hyperconnected, network society, posts on Twitter cause stock market crashes and overthrow governments (Pentland 2014). Unsurprisingly, life in social media is as messy as it is in the real world. Social media provides people exposure to new information and ideas, reflect their everyday highs and lows, make people engage in new friendships and break up old ones, make other people delighted or jealous by posting holiday and party photos, praise and complain about brands, and idolise the achievements of their descendants and pets. A bit simplified, users' behaviour in social media can be categorised into two types: rational/information seeking and emotional/affective seeking behaviours (Jansen et al. 2009). A desire to address a gap in information concerning events, organizations or issues is an example of information seeking behaviour in social media, whereas affective seeking behaviour stands for the expression of opinion about events, organizations or issues.

Studies have confirmed what every social media users already know: virtual public spheres attract users which Luoma-aho (2015) has named as hateholders and fakeholders. Social media provides hateholders continuously new targets and stimulus. Hateholders' behavior can be harsh, hurting and offensive, and it should therefore be condemned. Although the fighting against hateholders is not an easy task, it is possible because hateholders behavior is visible. Hateholders do not typically try to hide, quite contrary they pursue publicity. Fakeholders, on the other hand, act in the shadows. Although their behavior can also be harsh, hurting and offensive, it is difficult to get a hold on them. Acting through fake identities and using sophisticated persona management software, fakeholders aim to violate their targets.

Our goal is to investigate trolling phenomenon from social and psychological point-of-view. To facilitate analysis, sentiment analysis tool (Paavola & Jalonen 2015) is further developed to detect message automation which creates "noise" in the social media and makes it difficult to observe behavioral changes among human users. The work follows studies performed by Chu et al. (2012), Dickerson et al. (2014), and Clark et al. (2015) in which bot detection systems were devised. Components such as message timing behavior, spam detection, account

properties, and linguistic attributes were investigated. Variables were designed based on those components, and they were utilized in order to categorize message senders as humans, bots or cyborgs. The bot refers to a computer software that generates messages automatically, whereas cyborg in this context refers either to a bot-assisted human or to a human-assisted bot.

This paper is organized as follows: first, trolling behavior in the social media is analyzed from the psychological point-of-view. In the experimental part of the paper, automated bot detection is applied to Twitter messages. Finally, conclusions are drawn.

2. Trolling as a psychological phenomenon

Cambridge dictionary defines troll as "someone who leaves an intentionally annoying message on the Internet, in order to get attention or cause trouble" or as "a message that someone leaves on the Internet that is intended to annoy people" (Cambridge dictionaries online 2016). Oxford dictionary defines a troll as "a person who makes a deliberately offensive or provocative online post" or "a person who makes a deliberately offensive or provocative online post" (Oxford online dictionary 2016).

As a more scientific view, Claire Hardaker defines a troller in her paper as "a [computer-mediated communication] user who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement" (Hardaker 2010, 237). From a data of 186,470 social network posts, she identified four interrelated conditions related to trolling behaviour: aggression, deception, disruption, and success (Hardaker 2010, 225-236).

These definitions of trolling make no distinction between different motives behind deliberately offensive messages. These motives may be many-fold, ranging from emotional needs to rational ones, such as sidetracking an ongoing discussion by high jacking the topic.

Buckels, Trapnell and Paulhus (2014) studied the motivation behind trolling behaviour from a psychological viewpoint. They found out that self-reported enjoyment of trolling was positively correlated with three components of the Dark Tetriad, specifically sadism, psychopathy and Machiavellianism. The fourth component, narcissism, had a negative correlation with trolling enjoyment. To include the different aspects of trolling more comprehensively, Buckels et al also introduced a new scale, the Global Assessment of Internet Trolling (GAIT). Using GAIT scores they found sadism to have the most robust association with trolling behaviour, to the exceed that 'sadists tend to troll because they enjoy it'. (Buckels et al 2014).

Buckels et al (2014) define trolling as having no apparent instrumental purpose. One effect of succesful trolling, replacing a factual (or at least civilized) online discussion with a heated emonational debate with strongly emotional arguments, can, however, be used as a tool. The motivation behind this type of trolling behaviour is different from those who have an emotional need as basis.

All offensive texts, naturally, cannot be counted as trolling. What is seen as offensive by different cultures and individuals may vary, as Hardaker points out in her article. The final jugdement call, whether a post or a single comment is acceptable by the rules of the discussion forum, is typically made by a human moderator. The amount of data in large networks, however, makes it impossible for human moderators to follow all discussions taking place in real-time. To help moderators in this task, several types of methods have been implemented. One way is to utilize the users themselves. The easiest way is to provide the users with a way of alerting the moderator once content seen as inappropriate has been detected. The problems with this solution are that firstly, all the workload is carried by the moderator, and secondly, judgement is passed by a case by case principle, giving no credit to overall behaviour of a single user.

The implementation of different Trust and Reputation systems (TRSs), however, enable the users to rank their peers and allow building a more balanced ranking system. In one such example, F.J. Ortega, Troyano, Cruz, Vallejo, and Enriquez (2012) present their PolarityTrust system that they conclude to be a reliable system for assessing the trusworthiness of users in a network. Other TRS systems include algorithms such as

EigenTrust, Fans Minus Freaks, Signed Spectral Ranking, Negative Ranking, Non-Negative propagation approach and Action–Reaction approach. (Ortega et al 2012). The use of such systems help the moderator to have an overall peer view of a network user and thus helps to make a decision in a single case whether to ban a user or just give him or her a warning, for example.

Entirely different approach is to use algorithms that have the ability to spot undesired behavior and alert the moderator. One such example is provided by Moore, Nakano, Enomoto, and Suda (2012), who tested automatic labeling algorithms to identify cyberbullying. The labeling system had some success, but only as a tool for alerting a moderator. Moore et al (2012) suggest that the techniques described in their paper can also be used in relation to other problematic uses, just as trolling. This paper describes another attempt of this type of approach.

3. Tools for detection of message automation

Goals of the analyses are

- Analysis of how to detect fakeholders. The essential issue is to find out properties indicating that a message is sent by a bot or a cyborg. First step is to tag these kind of messages manually in order to use classification models on them. This procedure provides *ground truth* – the set of user accounts classified as bot or cyborg reliably by a human. As the number of analyzed user accounts has to be in the scale of several thousands, this part is time consuming.
- The sentiment analysis tool (Paavola & Jalonen 2015) is trained to detect message automation. This provides improved social media analytics tool.

In our paper, case study data consists of Twitter messages in Finnish language discussing Syrian refugee crisis. Twitter is a microblog service, where, on average, 500 million messages called tweets (140 characters in maximum) are posted on a daily basis. The openness of the Twitter platform allows, and actually promotes, automatic sending of messages. It is increasingly common to send automated messages to human users for attempting to influence them and to manipulate sentiment analyses (Clark et. al. 2015). Social media analyses can be skewed by bots that try to dilute legitimate public opinion.

Simple bot detection mechanisms analyze account activity and the related user network properties. Chu et al. (2012) utilized tweeting timing behavior, account properties and spam detection.

An example of a more advanced study is provided by Dickerson et. al. (2014). Their aim was to find out the most influential Twitter users on discussion about an election in India. To make this kind of assessment requires exclusion of bots from the analysis. Authors created a very complex model with tens of variables in order to decide whether the user is a human or a bot. Nineteen of those variables were sentiment based. Main findings were that bots flip-flop their sentiment less frequently than humans, humans express stronger sentiments, and humans tend to disagree more with the general sentiment of the discussion.

Sophisticated bot algorithms emulate human behavior and the bot detection must be performed from linguistic attributes (Clark et. al. 2015). In their work, authors used three linguistic variables to determine whether the user is a bot: the average URL count per tweet, the average pairwise lexical dissimilarity between user's tweets, and the word introduction rate decay parameter of the user for various proportions of time-ordered tweets. With these parameters, the authors were able to classify users as humans, bots, cyborgs or spammers. Authors concluded that for users, these four attributes are densely clustered, but can vary greatly for automated user accounts.

4. Case study

The development of our automatic bot detection systems was started with a cross-topic Twitter dataset, collected from September 17th, 2015 to September 24th, 2015. It covered more than 977 000 tweets in Finnish which were sent by more than 343 000 users. To develop an automatic classification system a ground truth dataset needs to be created. Among the collected data we randomly chose 2000 users, who had sent at least 10 tweets during that one week period. The sample set contains 83 937 tweets in total. We also extracted the profile data of those 2000 users from the Twitter. The tweets in other languages except Finnish were excluded from the dataset.

For each sampled user we used the following procedure. We carefully checked the text content of the tweets. We also checked other properties such as–tweeting application, the number of friends and followers, and in some cases, the user's homepage. The list of social media applications (such as Google, Facebook, and LinkedIn), mobile applications (such as Twitter for iPhone and Twitter for Android), and automation applications (such as dlvr.it and IFTTT) were build based on Internet searches. Lists contain also applications that were not found in the final refugee dataset. Lists were utilized in building features that were used in the classification.

In short, we labeled the user as a human or a bot based on the tweets' text, other information carried by tweets, the information contained in the user profile, and in some cases, external data. It took more than a minute on average to classify a user.

Application and the further development of the automatic classification system to the refugee related messages

The automatic bot detection system was further developed and applied to refugee-related messages in Twitter. The level of automation was used as classification to make the difference between bot and cyborg. If all messages sent by the user were interpreted as automated messages, the classification was ruled to be a bot. If only a part of messages seemed to originate from automation, the classification was ruled to be a cyborg. However, if less than quarter of messages were automated, the classification was a human. The collection of the refugee crisis tweets was based on Finnish key words and abbreviations. The free Twitter search API was used. The refugee dataset was collected from 6th of December 2015 to 3rd of February 2016. The whole dataset contained 59 491 tweets from 15 504 users. The set contained also some tweets in other languages, but those were excluded from the dataset based on the results of Language Detection Library (Shuyo 2016). After that, the Twitter users who had less than 10 tweets in the dataset were also excluded.

The final refugee dataset contains 31092 tweets from 855 Twitter users. Visualizations below show the most common hash tags as a word cloud, and locations where tweets have been posted. The latter data is available only if user has allowed geolocation data to be included.

The system was used to classify each user as either an automated user or as a human user. The automated user can be a cyborg or a bot. A weighted Random Forest algorithm was used. The refugee dataset was divided into a training set (684 users) and a test set (171 users), for us to be able to evaluate the performance of the automatic classifier. The case study was performed by using RStudio IDE version 0.99.442 on a 64-bit Ubuntu 14.04 LTS platform, and installed R version 3.2.2. The used random forest implementation package was Breiman and Cutler's Random Forests for Classification and Regression version 4.6-12.

Table 1: Confusion matrix showing very good accuracy and low number of false positive automatic classifications

		Classified by humans	
		Human	Bot or Cyborg
Classified by the pilot system	Human	137	6
	Bot or Cyborg	4	24

The test set confusion matrix of 171 Twitter users is presented in table 1. The recall is 80 percent, 24 out of 30 Twitter users that were manually classified by humans as bots or cyborgs were found as automated users in the test set by the pilot system. On the other hand, the precision is 86 percent, which indicates that most of the users the system classified as automated were manually classified as bots or cyborgs.

Some of the features the Random Forest algorithm found most important were the number of other users mentioned in the user's tweet on average, the number of links in the user's messages on average, and the type categories of the sending application (social media application, mobile application, or automation application). Some sentiment related features were also near the top of the list of about 30 features tried.

Figure 2 visualizes changes in the data set after automated messages were excluded. A notable difference can be seen in the most active tweet accounts. After excluding automated messages, active human participants in the discussion can be identified. In tweet locations no change is observed. Thus, automated accounts do not reveal this information. Only minor changes are seen in the hashtag list and sentiment result.

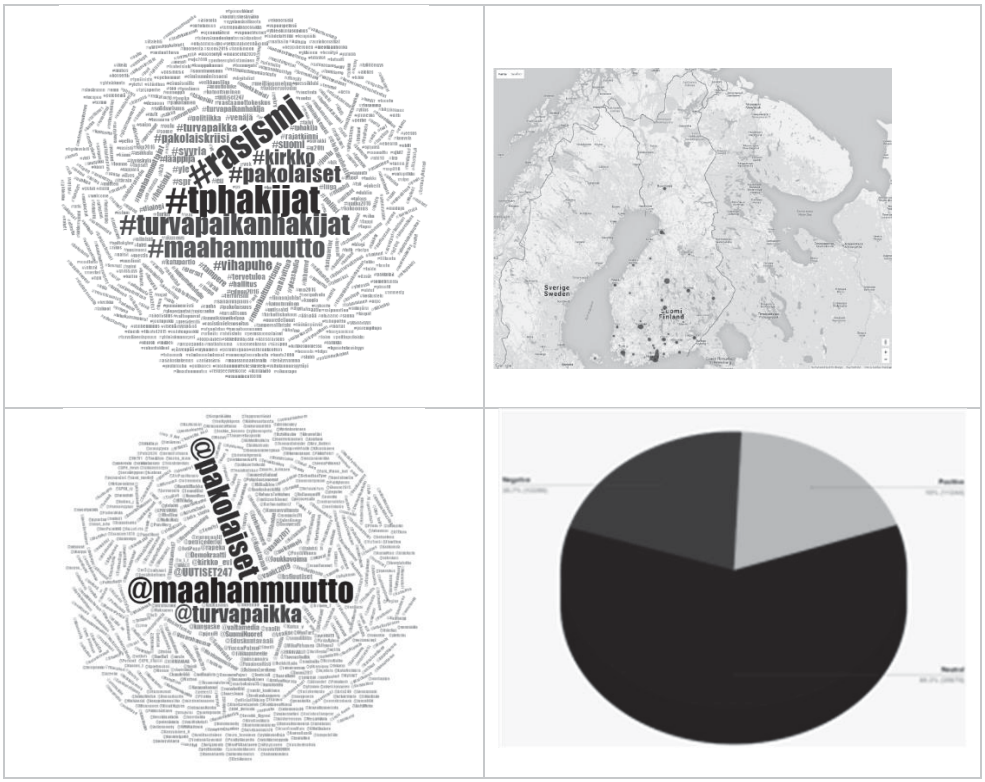


Figure 1: Visualizations of the collected data set. Top left: Wordcloud representation of the most commin hashtags (Translations: rasismi = racism; kirkko = church; pakolaiset = refugees; turvapaikanhakijat or tphakijat = asylum seekers; maahamuutto = immigration; vihapuhe = hate talk). Top right: Tweet locations. Bottom left: The most active tweet accounts. Bottom right: Sentiment analysis

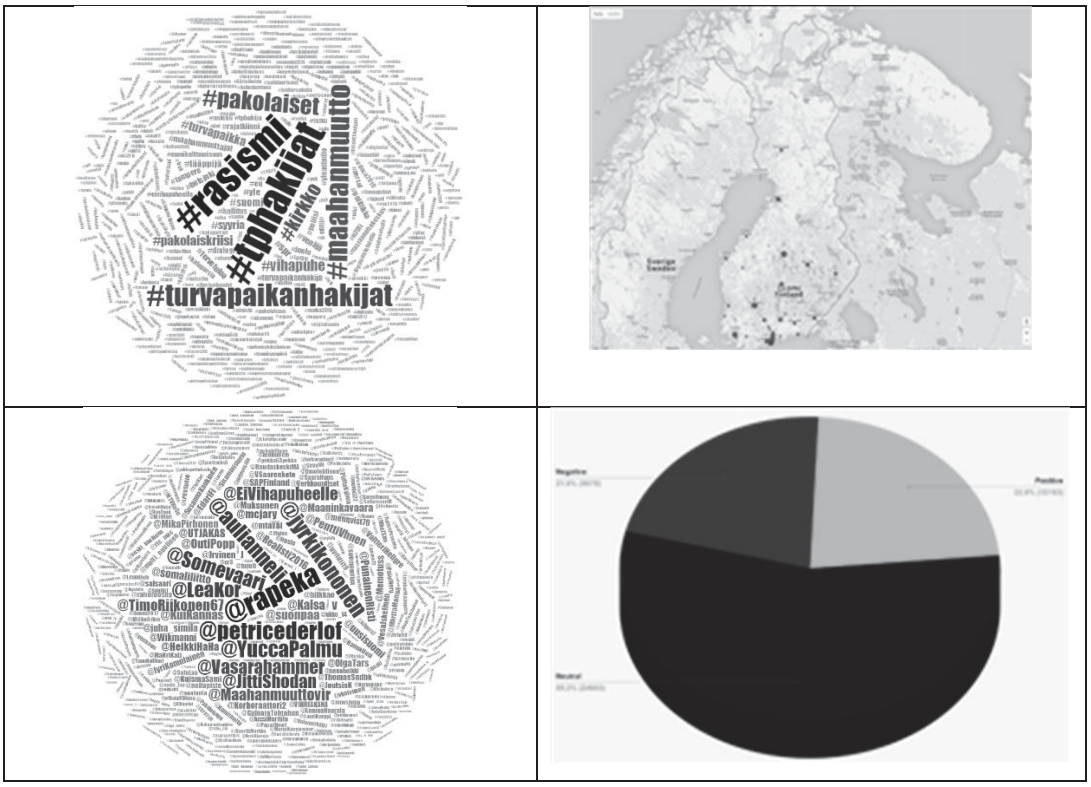


Figure 2: Visualizations of the collected data set. Top left: Wordcloud representation of the most commin hashtags (Translations: rasismi = racism; kirkko = church; pakolaiset = refugees; turvapaikanhakijat or tphakijat = asylum seekers; maahamuutto = immigration; vihapuhe = hate talk). Top right: Tweet locations. Bottom left: The most active tweet accounts. Bottom right: Sentiment analysis.

5. Discussion

In today's era of information overload, individuals and groups try to get their message across by using forceful language, engaging in dramatic (violent) actions, or by posting video clips or pictures on social media (Nacos et al. 2011, 48). The politically-driven mass media is partly behind this information overload on individuals. Aggressive behaviour is increasing in social media because of the technical ease with which trolling can be carried out. In social media all kinds of values become interwoven with each other.

During our work we found out that the definitions we used for trolling were not specific enough. Human communication, including malevolent one, is so diverse and contextual that in order to create automatic classification systems for trolling, a clearly defined behavioural pattern is needed. Trolling, as means of either distracting a conversation or simply provoking an emotional answer can utilize multiple context-based ways of influence. A positive word or sentence can, in the right context, actually be an insult. In order to enhance our algorithm we need to define the behavioural patterns we seek (such as trolling) in a way that can be translated into a more abstract form.

The one general psycho-sociological way to deal with trolls, who are systematically spamming information, is to limit our reactions to reminding others not to respond to them. When we live in a rhizome meshwork, the only course of action is not to feed the trolls. Trolling is a game of identity deception, in which the troll attempts to pass himself off as a legitimate participant, sharing the group's common interests and concerns. A troll can disrupt the discussion in a newsgroup, disseminate bad advice, and damage the feeling of trust in the newsgroup community.

The developed application was able to detect bots accurately, and the results were used to remove automated messages from the discussion. The case study work offers a solid base for further development. The main weakness of the pilot system is that it is quite heavily dependent on many Twitter specific features. It can also be the case that some of the used features are effective only when the bots and cyborgs are not trying to conceal themselves. We are constantly working towards extracting and creating new and more complex features, especially from the messages' text content. One example is to find out how much a single user's text content is changing from one message to another. This process is guided by the published research (Chu et al. 2012; Clark et al. 2015; Dickerson et al. 2014). The features based on the text content might have the advantage of being more easily applicable on the other social media forums outside Twitter, and more capable of finding concealed bots and cyborgs.

References

Bae, Y. & Lee, H. (2012) Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science & Technology*, 63(12), pp. 2521–2535.

Berthon, P. R., Pitt, L. F., Plangger, K. & Shapiro, D. (2012) Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons*, 55(3), pp. 261–271.

Buckels, E.E., Trapnell, P.D. and Paulhus, D.L. (2014) "Trolls just want to have fun", *Personality and Individual Differences* 67, pp. 97–102.

Cambridge dictionaries online (2016). http://dictionary.cambridge.org/dictionary/english/troll. Cited 13th of January 2016.

Chu Z., Gianvecchio, S., Haining, W., & Sushil, J. (2012): Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?, *IEEE Transactions on Dependable and Secure Computing*, Vol. 9, No. 6, November/December 2012.

Clark, E.M, Williams, J.R., Jones C.A., Galbraith, R.A., Danforth, C.M., Dodds, P.S. (2015) Sifting Robotic from Organix Text: A Natural Language Approach for Detecting Automation onTwitter, Journal of Computational Science (2015), http://dx.doi.org/10.1016/j.jocs.2015.11.002.

Coyne, R. (2014) The Net Effect. Design, the rhizome, and complex philosophy, online, http://www.casa.ucl.ac.uk/cupumecid_site/download/Coyne.pdf.

Deuze, M. (2011). Media Life. *Media, Culture and Society*, 33(1), pp. 137–148.

Deleuze, G. & Guattari, F. (1983). *On the Line*. Translated by John Johnston, Semiotext(s), New York.

Dickerson, J.P., Kagan, V., & Subrhamanian, V.S. (2014) Using Sentiment to Detect Bots in Twitter: Are Humans mire Opinionated than Bots?. *Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, August 2014.*

Habermas, J. (1989) *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*, MIT Press, Cambridge, MA.

Hardaker, C. (2010) "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions", *Journal of Politeness Research*, 6, pp. 215–242.

Jansen, B. J., Zhang, M., Sobel, K. & Chowdury, A. (2009) Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11), pp. 2169–2188.

Karppi, T. (2014) *Disconnect. Me – User Engagement and Facebook*, Doctoral Dissertations, Annales Universitatis Turkuensis, Ser. B Tom. 376, Humaniora, University of Turku.

Kaplan, A. M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53, pp. 59–68.

Lee, S. & Cude, B. J. (2012) Consumer complaint cannel choice in online and off-line purchases. *International Journal of Consumer Studies*, 36, pp. 90–96.

Luoma-aho, V. (2015) Understanding Stakeholder Engagement: Faith-holders, Hateholders & Fakeholders, *Research Journal of the Institute for Public Relations* Vol. 2, No. 1 (Winter, 2015).

Malgin, A. (2015) Kremlin Troll Army Shows Russia Isn't Charlie Hebdo. *The Moscow Times*, online, http://www.themoscowtimes.com/opinion/article/russia-is-not-charlie/514369.html.

Moore, M.J., Nakano, T., Enomoto, A. & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. Computers in Human Behavior 28, 861-867.

Nacos, B. L., Bloch-Elkon, Y., and Shapiro, R. Y. (2011). *Selling Fear. Counterterrorism, the Media, and Public Opinion*, The University of Chicago Press, Chicago.

Ortega, F.J., Troyano, J.A., Cruz, F.L., Vallejo, C.G. and Enriquez, F. (2012) "Propagation of trust and distrust for the detection of trolls in a social network", *Computer Networks* 56, pp. 2884–2895.

Oxford online dictionary. http://www.oxforddictionaries.com/definition/english/troll. 13JAN2016.

Paavola, J. and Jalonen, H. (2015) An Approach to Detect and Analyze the Impact of Biased Information Sources in the Social Media, Proc. *ECCWS* 2015.

Pentland, A. l. (2014) *Social Physics. How Good Ideas Spread – The Lessons from a New Science*, The Penguin Press, New York, NY.

Robertson, S. P., Douglas, S., Maruyma, M. & Semaan, B. (2013) Political discourse on social networking sites: Sentiment, in-group/out-group orientation and rationality. Information Polity: *The International Journal of Government & Democracy in the Information Age*, 18(2), pp. 107–126.

Shuyo N. (2016): https://github.com/shuyo. Cited 14th of February 2016.

Responding to North Korean Cyberattacks

Ji Min Park[1], Neil Rowe[2] and Maribel Cisneros[3]
[1]Republic of Korea Air Force, Seoul, Republic of Korea
[2]U.S. Naval Postgraduate School, Monterey, USA
[3]U.S. Army, Washington, USA
kimhkmin@naver.com
ncrowe@nps.edu
maribel.cisneros.mil@mail.mil

Abstract: North Korea has engaged in repeated cyberattacks on South Korea in the last ten years. These are consistent with their other provocations such as their nuclear weapons programs and aggressive activities along the border. The damage of these attacks is increasing, and they are a significant annoyance that is preventing progress on resolving the conflict between the two states. This paper will first survey the dramatic differences between North Korea and South Korea in cyberspace, and summarize the continuing evolution of North Korean cyberattacks. Attribution of most attacks to North Korea is not difficult because many of the same methods are used repeatedly, and many of the attacks are timed to coincide with important anniversaries. Thus, unlike with cybercriminals, these cyberattacks can often be predicted and recognized, and preparations can be made for them. North Korean cyberattacks often violate international law since they indiscriminately target civilians since the state considers itself not subject to international law. Nonetheless, there are many things South Korea can do to respond. International law can be invoked even without cooperation by the parties, and sanctions can be imposed. Defenses can be strengthened by better coordination of defenses between the private and governmental sectors through a unified early-warning defense. Repeated cyberattacks also justify a counterattack according to international law. North Korea has a limited set of state-mandated software on a small set of networks, and this makes widespread effects of cyberattacks on them quite possible. The main difficulty is getting access to these networks, but there are ways, especially given the growing use of smartphones in North Korea. As for attack goals, what North Korea fears the most is information about the rest of the world getting to their citizens, so sending accurate such information should be a major goal of cyber-counterattacks -- data destruction is unnecessary.

Keywords: North Korea, cyberattacks, response, attribution, defense, offense, international law

1. Introduction

The two Koreas are divided by ideological issues and have had a hostile relationship since 1950. Since the 2000s, North Korean provocations have shifted to cyberspace and to what they call “psychological operations” (Hewlett-Packard Security Research, 2014). These provocations attempt to express their political will and create a favorable environment for negotiation. Typically, North Korean official media first makes critical statements, then conducts cyberattacks, and then possibly conducts armed provocations. If South Korea could plan good responses to North Korean cyberattacks, North Korea might stop their provocations before they escalate. Good defensive and offensive countermeasures to cyber provocations can help reduce threats and prevent needless damage. However, it is important to find appropriate responses in cyberspace, a domain for which countries have limited experience (Lee et al, 2015; Park, 2015).

2. Cyberattacks on South Korea

2.1 Korean cyber capabilities

North Korea has separate domestic intranets for citizens, the Ministry of People’s Security, the Ministry of State Security, and the military. The unclassified intranet “Kwangmyong” connects 3,700 organizations with 50,000 estimated users. Most North Korean computers use the locally developed operating system Red Star, based on open-source Linux software and thus subject to the usual vulnerabilities of Linux (Lee, 2011). To connect to the Internet, optical-fiber cables link from China and use Chinese IP addresses. Use of the Internet in North Korea is strictly controlled by the North Korean government, and users are estimated to be only hundreds of high-ranked officials. Furthermore, operating hours are limited because an electrical power shortage. It was significant progress when they opened six official Web sites to the public. Cellular networks provide 3G-network cellphone service, and the number of users is estimated at 2.5 million, 10% of the population. Though general users cannot access the Internet, it can be reached from the cellular network (Mansourov, 2014).

Attacks directly from North Korea are easy to attribute because they have few IP addresses (Lim et al., 2013). So North Korea usually launches attacks from elsewhere, especially China. Since there are no voluntary hacker groups in North Korea because the government strongly controls all networks, all cyberattacks coming directly or indirectly from North Korea can be attributed to the North Korean government (HP Security Research, 2014). However, because of economic sanctions, North Korea has had difficulty in acquiring technologies and devices for developing their cyber capabilities.

On the other hand, South Korea has the 12th highest number of Internet users in the world, and depends considerably on the Internet, maintaining high Internet availability (MSIP and KISA, 2014). In South Korea, 98% of businesses with more than 10 employees are connected to the Internet, and 86.7% of employees are using the Internet for business. In addition, most citizens have smartphones using 4G technology. South Korea is connected to the Internet via a 27 Tbps connection through undersea cables and communication satellites. In 2014, South Korea had 122 million IPv4 addresses, which is the 6th highest rank in the world, and had 1018 Autonomous System (AS) numbers as the 13th rank. The national Domain Name Service (DNS) consists of 15 sites, and its average daily query total is about 1.6 billion. Internet services are provided by a number of private Internet Service Providers. South Korea uses many applications popular around the world such as Microsoft Office or Adobe Photoshop, but many South Koreans prefer to use domestic word processors and domestic antivirus software. Furthermore, due the policy of favoring domestic software manufacturers, government organizations tend to use domestic products.

South Korea's large cyber resources and dependency on digital culture are a disadvantage in cyberwarfare (Kshetri, 2014) since so many potential targets exist in its extensive cyberspace (Lim et al., 2013). In addition, it is hard for the South Korean government to control its country's cyberspace because much of it is in the private sector, and role division between organizations is not always settled (Boo, 2013). South Korea also provides diverse electronic-government services which are a tempting target in cyberwarfare because they have much personal information.

2.2 Cyberattacks from North Korea against South Korea

North Korea has said it is preparing for "fourth generation warfare" by establishing cyberwarfare units and developing capabilities (IUE, 2014). More than 70,000 cyberattacks have been conducted against South Korean government and civilian Web sites, and many of these have been attributed to North Korea (Chae, 2013). Some of the most serious attacks have been:

- On June 10, 2004, South Korea government sites were attacked by malware that appeared to be coming from China (Kim, 2010). According to statistics, among 301 damaged computers, 222 devices belonged to the government and 79 to private companies and universities. Based on the analysis, the origin of the attack was China, but the IP addresses were Chinese ones being leased by North Korea. Secret information related to national security was leaked for six months (Chae, 2013). In addition, North Korea accessed South Korean military wireless-communication networks (HP Security Research, 2014).
- North Korea penetrated South Korea's military communication channel in 2005 during Ulchi Focus Lens, the annual combined military exercise with the United States (Ventre, 2011).
- The U.S. State Department was attacked by unknown entities in cyberspace in June 2006 (HP Security Research, 2014). Although detailed information is not available, the South Korean military reported that North Korea's Unit 121 was implicated in this attack.
- In March 2007, the Korean Army and the National Institute of the Environment were attacked (Mansourov, 2014). After the hackers obtained certificate passwords, they stole information related to chemical-accident response. South Korea's government announced that the malware was from a foreign country, and it could have been from North Korea (Lee, 2009).
- North Korea in 2008 sent malicious emails with Trojan Horses to the South Korean military, and social engineering attempts such as spear phishing were also identified (Ventre, 2011).
- July 2009 saw a distributed denial-of-service (DDoS) attack targeting 21 Web sites, including government sites, media, and financial institutions. Hackers used sophisticated methods such as automatic deletion of source files and destruction of zombie hard disks to hide evidence of the attacker's identity (Chae, 2013). The attack exploited over 400 servers in the world to make tracing hard. The total number of bots was

approximately 20,000 devices. Among them, 12,000 infected computers were located in South Korea, and others were in foreign countries (Mansourov, 2014).

- In 2010, several cyberattacks for information leakage were targeted against South Korean military officers via malware (Ventre, 2011). In July, there were DDoS attacks against several government sites (Mansourov, 2014) conducted by the same botnets used in 2009 and the signatures of the malware were the same (Jung, 2010).
- In January 2011, Free North Korea Radio was attacked by North Korea. Unusually, this attack came from North Korea directly, which indicated that they wanted the attack to be attributed.
- In March 2011, a large DDoS attack was conducted against the South Korean government and private services. Over 700 servers and 100,000 infected PCs were mobilized (Chae, 2013). The attack was similar to previous attacks known to be from North Korea (Ahn, 2013) and used illegal game sites to spread malware that infected PCs and set up botnets (Mansourov, 2014).
- In April 2011, the South Korean bank Nonghyeop was hacked and their banking service was paralyzed for several days (Chae, 2013). The South Korean prosecutor said that this attack was conducted by North Korea's RGB after seven months of preparation (Park, 2011).
- In June 2012, North Korea attacked the Web site of Joongang Ilbo, a conservative press company (Mansourov, 2014). The South Korean Cyber Terror Response Team in the National Police Agency reported that the IP addresses of this attack were related to the North Korean Ministry of Posts and Telecommunications (Park, 2013). Databases related to news articles and pictures were destroyed by this attack.
- In March 2013, North Korea attacked PCs, servers, and automated teller machines of six broadcast companies and financial institutes, causing deletion of data and service interruptions (Chae, 2013). This attack was an advanced-persistent-threat attack similar to the Nonghyeop hacking in 2011 (Boo, 2013). This attack was prepared over eight months, and approximately 57,000 PCs and servers were damaged (Lee, 2015). IP addresses belonging to North Korea were identified 13 times, and many intermediate pathways used were identical to their past attacks. In addition, 30 types of malware among the total 76 instances of malware used for this attack were used in previous North Korean attacks (HP Security Research, 2014).
- In April 2013, the international hacking group Anonymous attacked North Korea's networks (HP Security Research, 2014). Apparently in response, malicious smartphone applications spread as free games infected approximately 20,000 devices in South Korea from May to September (Mansourov, 2014).
- In June 2013, a DDoS attack was attributed to the DarkSeoul hacking group related to North Korea's Lab 110 (HP Security Research, 2014). The malware used stolen passwords and destroyed hard drives (Lee, 2015).
- In 2014, Sony Pictures Entertainment was attacked apparently in response to the film *The Interview*, which depicted the assassination of a North Korean leader (HP Security Research, 2014). The US FBI attributed this cyberattack to North Korea and concluded that the attack was malware that both stole information and destroyed data (FBI, 2014). Experts have estimated the damages were up to $100 million (Richwine, 2014).
- In December 2014, Korea Hydro and Nuclear Power Co., Ltd (KHNP) was hacked, and the South Korean Public Prosecutor's Office claimed North Korea as the source based on traces of IP addresses that belonged to North Korea (Kim, 2015).

Figure 1 shows the monthly trends in cyberattacks 2004-2015. Many attacks occurred in March, April, and July. In March there is an annual large-scale combined military exercise involving South Korea and the United States; apparently cyberattacks are intended to protest and disturb the exercises. June 25 is the anniversary of the start of the Korean War, and July 4 is the U.S. Independence Day. These coincidences suggest that North Korea is involved in attacks in those months.

From 2004 to 2008, most attacks were information-gathering from government and research agencies. Since 2009 there have also been DDoS attacks against the private sector such as media and financial institutes. Hackers have increasingly targeted groupware and antivirus systems, and damages have increased. In addition, in December 2014, they hacked the KHNP, part of the national infrastructure. So the targets of North Korea's cyberattacks have shifted from intelligence collection to DDoS attacks, and then to advanced persistent threats on the private sector. The increasing scope of these attacks means that a lethal cyberattack from North Korea is increasingly possible.

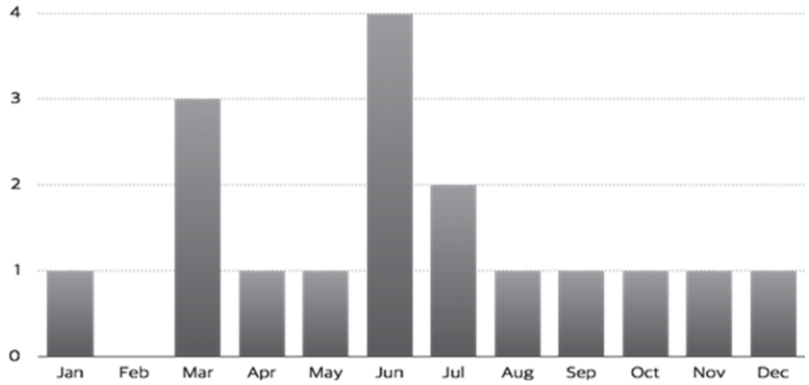


Figure 1: Monthly statistics of North Korea's major cyberattacks from 2004 to 2015

2.3 Attribution of the cyberattacks

How do we know that the abovementioned attacks were from North Korea? Attribution can be done by many methods (Rowe, 2015, "Attribution"). Data analysis that indicates similarities between attacks is one way. For instance, (Seo, Won, and Hong, 2011) compared the July 2009 and March 2011 attacks by inspecting the traffic initiated from infected systems at the Pohang University of Science and Technology. They found that both attacks (1) were autonomous with a predefined target list and start date and time; (2) used botnets with low-rate (54.2kbs) attacks to remain undetected on the infected hosts; (3) used multiple forms of DDoS attacks (TCP SYN floods, ICMP floods, UDP floods, and HTTP GET/POST flood); and (4) included instructions on the bots to delete documents and corrupt the Master Boot Record. These similarities are beyond chance.

The U.S.-Computer Emergency Response Team (U.S. CERT, 2013) determined that the code in the March 2013 attacks was designed to avoid South Korean antivirus signatures similarly to the two abovementioned attacks. FireEye analysts corroborate the U.S.-CERT assessments for this particular attack (Pidathala, Khalid, Singh, and Vashisht, 2013) and found that the malware could also disable the popular AhnLab South Korean antivirus software by issuing a taskkill command for it. The code was to be "dropped" by another program, which could have been over HTTP as there was an increase of executable HTTP downloads a month before the attacks (Dell SecureWorks, 2013). This basic attack method is mentioned in the World War C report about North Korea: hacking websites with malware to take over their operating system and disable their antivirus software (Geers et al, 2013). McAfee assessed that the same group has been behind all the cyberattacks against South Korea since 2009. (Sherstobitoff and Liba, 2013) concluded these attacks to be related because it was mostly the same code using the same password for the zip files, with some added functionality such as file extraction in later versions. Symantec has also linked cyberattacks against South Korea since 2009 to include the June 2013 attack (Symantec, 2013).

It is also clear that South Korea has been the primary target of these attacks because of their dates, their attempts to evade specific South Korean antivirus software, and their target lists of South Korean organizations and Korean-military search terms. The autonomous nature of the attacks also points to North Korea which has limited Internet connections for command-and-control. Most of the infected bots were in South Korea (Seo et al, 2011), which suggests North Korea's involvement since they have repeatedly attempted to grow botnets in South Korea by infected games, phishing attacks, and HTTP executables. The code also tried to extract military files of special interest to North Korea. Moreover, the increased level of malware sophistication and added functionality with each attack are consistent with organized military cyber units that continue to develop their skills.

3. Possible responses to North Korean cyberattacks

3.1 Legal responses

Although North Korea has been termed a "rogue state" because it often ignores international law, the international community can affect it through sanctions (Cisneros, 2015). South Korea can invoke Articles 39, 41, and 42 of the U.N. Charter (1945), present collected evidence to the U.N. Security Council, and ask them to take action against North Korea. Under Article 39, the Council will evaluate the evidence and determine if "the existence of any threat to the peace exits, breach of the peace, or act of aggression" and "what measures shall

be taken in accordance with Article 41 and 42." The Council will attempt to settle the dispute without the use of armed force, and if that fails to stop the threat, it could authorize South Korea to use military operations to maintain or restore peace.

Responding to cyberattacks involves a significant amount of time and resources, and requires response personnel to have technical knowledge. South Korea can sue North Korea for damages through the International Court of Justice (ICJ) which enforces customary international law (ICJ, 2015). The ICJ can then review the evidence, conduct investigations, and issue their ruling, which North Korea, as a signatory of the U.N. Charter, is mandated to abide by (ICJ, 2015). Even a single attack can be sufficiently damaging that the perpetrators can be held accountable for monetary damages. That is, the ICJ can enforce tort law, which stipulates that if an aggressor conducts an intentional act to cause harm, the victim can pursue a lawsuit to obtain compensation; cyberspace torts are classified as intentional torts (Grama, 2010).

North and South Korea are still legally at war so cyberattacks between them also need to abide by the law of armed conflict. The basic principles are distinction, proportionality, military necessity, and humanity (Carr, 2011). Since North Korean systems are tightly controlled by their government, targeted cyber-counterattacks on North Korean computer systems will be a direct attack on their military and government systems, which will abide by the principle of distinction of only targeting combatants. The counterattacks also can easily meet the proportionality criteria, as there is not much to attack. The counterattacks can also meet the military-necessity criteria, as the cyberattack will seek to only "weaken the military forces of the enemy" (ICRC, 2002). Additionally, such attacks can be consistent with the humanity principle because they can avoid unnecessary suffering and injuries by attacking only cyberspace.

Since we have strong evidence that North Korea is responsible for many attacks on South Korea, the U.N., through the ICJ, can subpoena packet information between these two states to facilitate accountability after a cyberattack. The U.N. will need international cooperation to obtain information from the different traversed routers. The broad perspective of the U.N. can take into account non-cyber issues such as the possibility of war in the Korean Peninsula, maintaining peace talks, handling North Korea's established military objective to reunite the Koreas, and addressing North Korea's inability to feed their citizens without international aid (Worden, 2008). Previous cases have shown that if state-sponsored attacks continue to go unpunished, civilian organizations will continue to be targeted.

3.2 Defensive countermeasures

A defender is not powerless in cyberspace (Singer and Friedman, 2014). Cyber-defensive countermeasures are possible based on the idea of active defense (Harrington, 2014) and these can be used by South Korea. Beaconing transmits current user information, such as IP addresses, when the stolen file is opened to enable tracing of the theft. Honeypots are systems designed solely to collect information about attacks (Harrington, 2014) and can be effective against distributed denial of service, malware installation for advanced persistent threats, and other kinds of attacks. Sinkholing intercepts malicious traffic from botnet clients by masquerading as one of its command-and-control servers (Harrington, 2014), and can thereby foil botnets.

Attackers can choose a type, time, and target of attack, but they must subvert all protective layers to make their attack successful. This means a cyber early-warning system can work well and provide a wider viewpoint to defense since it only need recognize some clues to an attack (Robinson, Jones, and Janicke, 2015). A good cyber early-warning system should provide information about the current situation, the attacker, the targets, and the attack methods (Golling and Stelte, 2011). Cyberattacks motivated by political issues usually have five phases: latent tension, cyber reconnaissance, an initiating event, cyber mobilization, and a cyberattack (Carr, 2011). That suggests that defense should start collecting a good deal of cyber data in the early phases. Early information about subject, target, and method could be incorrect, but immediacy is more important than accuracy because it aids a quick defense. An early-warning system should also share threat information quickly between systems, and it is especially important to get warnings to critical infrastructure such as power plants and water-supply plants. Furthermore, since North Korea has exploited antivirus software to spread malware, cooperation with foreign cyber-security companies is also important.

3.3 Offensive countermeasures

South Korea can also try to counterattack North Korea since that is justified by the law of armed conflict as discussed above. Despite the many air-gapped systems in North Korea, such systems are hard to update and can have persistent vulnerabilities. Brittleness of enemy systems can be demonstrated by the ability to hack into enemy systems at any time as a kind of "retaliatory hacking" (Libicki, 2013; Harrington, 2014). Counterattacks can be quite limited, as a form of coercion (Flemming and Rowe, 2015) or a show of force to encourage North Korea to cease cyberattacks and offer compensation. Joint coercion with conventional military operations can also enhance bargaining power and prevent escalation.

When attribution of a cyberattack is easy to prove, South Korea could consider targeting the originating sites of the cyberattack as a simple way to satisfy the principle of proportionality (Rowe, 2015, "Distinctive"). Second, because the North Korea government strongly controls information about the outside world in violation of human rights (IUE and Ministry of Unification, 2014), their propaganda services make good targets. A good counterattack need not harm North Korea at all, just provide accurate information about the outside world on their intranets, since the North Korean government fears this information so much. Distributed denial-of-service attacks based on botnets are also possible against North Korea (Radunovic, 2013). The estimated cost of a distributed denial of service botnet that could attack national-level targets is only 6,000 euros.

Collateral damage, difficulties of damage localization of cyber weapons, and direct damage from cyberattacks could be redeemed to a certain degree by using reversible counterattacks (Rowe, 2010). Reversibility can be achieved using encryption attacks, obfuscating attacks, withholding-information attacks, or resource-deception attacks (Rowe, 2010). Damage cannot always be recovered fully in some cases such as time-sensitive operations, and reversibility is less possible as time goes on and adversaries attempt to stop the counterattack. But if attackers provide recovery or assume responsibility for most of the damages, criticism of the counterattack could be reduced.

One concern with counterattacks is the possibility of escalation of the conflict. But considering North Korea's limited Internet infrastructure, North Korea can only accomplish attacks if they are planned well in advance. If South Korea counterattacks it, the North Korean government should not be able to counter-counterattack quickly, particularly if the counterattack targets are intended for supporting attacks. And attacks on the North Korean government's propaganda services should not affect the North Korean public. Thus the risk of escalation with direct counterattacks is relatively less than those of viruses, worms, or other methods with hard-to-predict results.

4. Conclusion

The extensive infrastructure and diversity of South Korea makes cyberattacks an appealing option for North Korea. The last ten years have seen many such attacks, most of which were easy to attribute. But South Korea has many options in response. More legal measures are possible. A better coordinated defensive strategy is important, combining government and the private sector. Several kinds of offensive measures could be useful, and they do not need to destroy anything to be powerful.

References

Ahn, Y. (2013) "Study of Development Plan for National Defense System against Cyberattacks," *Review of the Korea Institute of Information Security and Cryptology*, Vol. 23, No. 2, pp. 48–54 (in Korean).

Boo, H. (2013) "Issue of Cyber Security and Policy Directions: Discussions for the Establishment of the Defense Ministry's Cyber Policy," *Journal of National Defense Studies*, Vol. 56, No. 2, pp. 97–122 (in Korean).

Carr, J. (2011) *Inside Cyberwarfare (2nd ed.)*, O'Reilly Media, Sebastopol, CA, US.

Chae, J. (2013) "The Changing Security Environment and Cyber Security," *The Journal of Political Science and Communication*, Vol. 16, No. 2, pp. 171–193 (in Korean).

Cisneros, M. (2015) *Cyber-Warfare: Jus Post Bellum*, Master's thesis, Naval Postgraduate School, Monterey, CA, US, March.

Dell SecureWorks (2013, March 21) "Wiper Malware Analysis Attacking Korean Financial Sector, retrieved from www.secureworks.com/cyber-threat-intelligence/threats/wiper-malware-analysis-attacking-korean-financial-sector/.

Federal Bureau of Investigation (FBI) National Press Office (2014) "Update on Sony Investigation," retrieved from www.fbi.gov/news/pressrel/press-releases/update-on-sony-investigation?utm_campaign=email-Immediate&utm_medium=email&utm_source=national-press-releases&utm_content=386194.

Flemming, D, and Rowe, N. (2015) "Cyber Coercion: Cyber Operations Short of Cyberwar," Proceedings of the 10th International Conference on Cyberwarfare and Security ICCWS-2015, Skukuza, South Africa, March, pp. 95–101.

Geers, K., Kindlund, D., Moran, N. and Rachwald, R. (2013) "World War C: Understanding Nation-State Motives behind Today's Advanced Cyberattacks," retrieved from www.fireeye.com/resources/pdfs/fireeye-wwc-report.pdf.

Golling, M., and Stelte, B. (2011) "Requirements for a Future EWS - Cyber Defence in the Internet of the Future," Proceedings of the 3rd International Conference on Cyber Conflict, Tallinn, Estonia, pp. 1-16.

Grama, J. (2010) *Legal Issues in Information Security,* Jones and Bartlett Learning, Sudbury, MA, US.

Harrington, S. (2014) "Cyber Security Active Defense: Playing with Fire or Sound Risk Management?," *Richmond Journal of Law and Technology*, Vol. 20, No. 4, pp. 1-41.

Hewlett-Packard (HP) Security Research (2014) *Profiling an Enigma: The Mystery of North Korea's Cyber Threat Landscape (HP Security Briefing Episode 16)*. Retrieved from community.hpe.com/hpeb/attachments/ hpeb/off-by-on-software-security-blog/388/2/HPSR%20SecurityBriefing_Episode16_NorthKorea.pdf

Hong, S. (2011) "North Korea's Cyberattack Methods, Advanced and Intelligent," *The Unified Korea*, Vol. 328, pp. 34–35 (in Korean).

Institute of Unification Education (IUE) and Ministry of Unification, South Korea (2014) *Understanding North Korea 2014*, Nuel-Pum Plus, Seoul, South Korea.

International Committee of the Red Cross (ICRC) (2012) "International Humanitarian Law: Answers to Your Questions," retrieved from www.redcross.org/images/MEDIA_CustomProductCatalog/ m22303661_IHL-FAQ.pdf.

International Court of Justice (ICJ) (2015) "Basic Documents: Statue of the Court," retrieved from http://www.icj-cij.org/documents/?p1=4&p2=2.

Jung, Y. (2010) "Recurrence of DDoS Attacks against the Blue House, the Ministry of Foreign Affairs, and Naver.com," retrieved from newshankuk.com/news/www.newshankuk.com/news content.asp? news_idx=20100708092500n5131 (in Korean).

Kim, H. (2010) "North Korea's Cyber Terror and Information Warfare Capabilities, and Cyber Security Countermeasure Proposals," retrieved from www.boan.com/news/ articleView.html?idxno=1391 (in Korean).

Kim, Y. (2015) "KHNP Hacking is Attributed to North Korea," retrieved from www.pressian.com/news/ article.html?no=124755 (in Korean).

Kshetri, N. (2014) "Cyberwarfare in the Korean Peninsula: Asymmetries and Strategic Responses," *East Asia*, Vol. 31, pp. 183–201.

Lee, M. (2011) "North Korean OS 'Red Star 2.0' Very Vulnerable against Cyberattacks," *Digital Daily*, retrieved from www.ddaily.co.kr/news/article.html?no=84158 (in Korean).

Lee, S. (2009). National important information was leaked through a hole of the military Internet. Yonhap News. Retrieved from news.naver.com/main/ read.nhn?mode=LSD&mid=sec&sid1=100 &oid=001 &aid=0002922925 (in Korean).

Lee, Y., Kwon, H., Lee, J., and Shin, D. (2015) "Development of Countermeasures against North Korean Cyberterrorism through Research Case Studies," *The Korean Journal of Defense Analysis,* Vol. 27, No. 1, pp. 71–86 (in Korean).

Libicki, M. (2013) "Brandishing Cyberattack Capabilities," retrieved from www.rand.org/content/dam/rand/pubs/research_reports/RR100/RR175/RAND_RR175.pdf.

Lim, J., Kwan, Y., Chang, K., and Baek, S. (2013) "North Korea's Cyber War Capability and South Korea's National Counterstrategy," *The Quarterly Journal of Defense Study Policy Studies*, Vol. 29, No. 4, pp 9–45 (in Korean).

Mansourov, A. (2014) *North Korea's Cyberwarfare and Challenges for the U.S.-ROK Alliance,* Korea Economic Institute of America Academic Paper Series 2014, Korea Economic Institute of America, Washington, DC, US, pp. 1–17.

Ministry of Science, ICT, and Future Planning (MSIP), and Korea Internet and Security Agency (KISA) (2014) *Korea Internet White Paper 2014*, GPRN 11-B551505-000008-10, Myeong-jin C&P, Seoul, South Korea.

Park, D. (2011) "Study of Hacking in Terms of National Cyber Security Policy," *Review of the Korea Institute of Information Security and Cryptology*, Vol. 21, No. 6, pp. 24–41 (in Korean).

Park, J. (2015) *Finding Effective Responses against Cyber Attacks for Divided Nations,* M.S. thesis, U.S. Naval Postgraduate School, Monterey, CA, US, December.

Park, Y. (2013) "Police Announced that Hacking on Joongang Ilbo in 2012 Was from North Korea," Yonhap News, retrieved from www.yonhapnews.co.kr/society/2013/01/16/0701000000AKR2013011 6090400004.HTML (in Korean).

Pidathala, V., Khalid, Y., Singh, A., and Vashisht, S. (2013, March 21) "More Insights on the Recent Korean Cyberattacks (Trojan.Hastati)," retrieved from www.fireeye.com/blog/technical/ botnet-activities-research/2013/03/more-insights-on-the-recent-korean-cyber-attacks-trojan-hastati.html

Radunovic, V. (2013) "DDoS – Available Weapon of Mass Disruption," Proceedings of the 2013 21st Telecommunications Forum (TELFOR), Geneva, Switzerland.

Richwine, L. (2014) "Cyberattack Could Cost Sony Studio As Much As 100 Million," retrieved from www.reuters.com/article/2014/12/09/us-sony-cybersecurity-costs-idUSKBN0JN2L020141209.

Rowe, N. (2010) "Towards Reversible Cyberattacks," Proceedings of the 9th European Conference on Information Warfare and Security, Thessalonika, Greece.

Rowe, N. (2015) "Distinctive Ethical Challenges of Cyberweapons," in Tsagourias, N., and Buchan, R., (eds.), *Research Handbook on Cyber Space and International Law*, Edward Elgar, Cheltenham, UK, pp. 307-325.

Rowe, N. (2015) "Attribution of Cyberwarfare," Chapter 3 in Green, J. (ed.), *Cyber Warfare: A Multidisciplinary Analysis*, Routledge, Oxon, UK, pp. 61-72.

Seo, S., Won, Y., and Hong, J. (2011, October) "Witnessing Distributed Denial-of-Service Traffic from an Attacker's Network," 7th International Conference on Network and Service Management (CNSM), Paris, France.

Sherstobitoff, R. and Liba, I. (2013) "Dissecting Operation Troy: Cyberespionage in South Korea," McAfee Labs White Paper, retrieved from http://www.mcafee.com/us/resources/white-papers/wp-dissecting-operation-troy.pdf.

Singer, P., and Friedman, A. (2014) Cybersecurity and Cyberwar: What Everyone Needs to Know (1st ed.), Oxford University Press, New York, NY.

U.S. CERT (2013, April 2) "South Korean Malware Attack," retrieved from www.us-cert.gov/sites/default/files/publications/South%20Korean%20Malware%20Attack_1.pdf.

Ventre, D. (2011) *Cyberwar and Information Warfare*, John Wiley and Sons, Hoboken, NJ, US.

Worden, R. (2008) *North Korea: A country study*, Library of Congress, Federal Research Division, retrieved from lcweb2.loc.gov/frd/cs/pdf/CS_North-Korea.pdf.

An Overview of Linux Container Based Network Emulation

Schalk Peach[1, 2], Barry Irwin[2] and Renier van Heerden[1]
[1]Council for Scientific and Industrial Research, South Africa
[2]Rhodes University, Grahamstown, South Africa
speach@csir.co.za
b.irwin@rhodes.ac.za
rvheerden@csir.co.za

Abstract: The objective of this paper is to assess the current state of Container-Based Emulator implementations on the Linux platform. Through a narrative overview, a selection of open source Container-Based emulators are analysed to collect information regarding the technologies used to construct them to assess the current state of this emerging technology. Container-Based Emulators allows the creation of small emulated networks on commodity hardware through the use of kernel level virtualization techniques, also referred to as containerisation. Container-Based Emulators act as a management tool to control containers and the applications that execute within them. The ability of Container Based Emulators to create repeatable and controllable test networks makes it ideal for use as training and experimentation tools in the information security and network management fields. Due to the ease of use and low hardware requirements, the tools present a low cost alternative to other forms of network experimentation platforms. Through a review of current literature and source code, the current state of Container-Based Emulators is assessed. The primary source of information is publications by the creators of the selected Container-Based Emulators. Each Container-Based Emulator is introduced with a brief summary of the history and requirements that lead to its creation. The reader is presented with a comparison of the specific kernel level virtualization technologies used to implement the virtualization sub-system of the Container-Based Emulator. The structural design of Container-Based Emulators is analysed and summarized to provide a concise view of the capabilities that users are presented with. An architectural model is introduced that can assist with the selection of a Container-Based Emulator, based on the requirements of the end user. The paper concludes with a summary of the current state of Container-Based Emulators, and proposes future research areas to be explored.

Keywords: Linux, containerisation, network emulation, container-based emulator

1. Introduction

Operating system level virtualisation, also known as containers, is applied in various circumstances such as application packaging and low overhead virtual private servers (Merkel, 2014; Soltesz, Pötzl, Fiuczynski, Bavier, and Peterson, 2007). When combined, Linux containers and network virtualisation techniques such as Linux bridges provides the base technologies to create low overhead virtual networks on commodity hardware (Bhatia, Motiwala, Muhlbauer, Mundada, Valancius, Bavier, Feamster, Peterson, and Rexford, 2008; Lantz, Heller, and McKeown, 2010). These software applications are often referred to as Container Based Emulators (CBE). CBEs provide a low cost, low maintenance alternative to other forms of network experimentation platforms. These systems can be deployed as sandboxed training and experimentation platforms for various fields including network administration and information security. The aim of this paper is to collect the current knowledge regarding software applications specialising in providing a frontend for the creation of the virtualised computer networks. A narrative overview methodology is used to assess the current state of open source Linux container based network emulation through a study of literature published on the subject matter. The reader is presented with the information required to select a CBE that is suitable to the task at hand.

1.1 Related work

Linkletter, 2015 provides reviews of most of the Container-Based Emulators discussed in this paper. The reviews provided are however focused on installation and usage. In contrast, this paper focuses on technical aspects of Container-Based Emulators. In (Pizzonia and Rimondini, 2014), a collection of network simulation and emulation platforms are compared based on the authors requirements for deployment in an educational context. The comparison subsequently focuses on deployment, ease of use and technology support, in preference to technical aspects of these systems. Salopek (Salopek, Vasic, Zec, Mikuc, Vasarevic, and Koncar, 2014) compares the overhead incurred in container-based emulators based on the method of node virtualisation.

1.2 Structure

In Section 2 a brief introduction to different types of network experimentation platforms is given, highlighting the advantages and disadvantages of CBEs. In Section 3 current open source CBEs are introduced. In Section 4,

the selected CBEs are compared at technology and feature level. Finally, in Section 5, conclusions on the current state of CBEs are presented, as well as highlighting alternative methods used to create network experiments using operating system level virtualisation.

2. Network experimentation platforms

Network experimentation platforms enable researchers to execute repeatable experiments, ensuring consistent results. These platforms can be constructed using various techniques. This section considers four broad technology areas that can be employed to create network experimentation platforms and gives a brief overview of each type.

2.1 Testbeds

A testbed is a deployment of computer and networking hardware that aims to replicate the conditions in which a software application will be utilised. Key goals of a testbed is to recreate expected conditions at the highest possible level of fidelity. One of the advantages of a testbed is absolute fidelity of the components used to construct the experimentation platform, resulting in repeatable tests. Due to the large number of hardware components and space required, the cost of deploying a testbed can be prohibitive.

2.2 Virtualisation

The cost of deploying a testbed can be reduced through the use of virtualisation. The functional fidelity of virtualised computer hardware is near perfect, complementing the repeatability of experiments. By replacing costly end user hardware with virtualised instances, the total hardware required is reduced. Additional advantages of virtualisation are reductions in physical space and maintenance requirements.

2.3 Simulation

An alternative to testbeds, with or without virtualisation, is simulation. Through simulation, hundreds to thousands of nodes in a computer network can be simulated on a single high end server. The behaviour of simulated components is restricted to the accuracy of the models employed to simulate the component. This could result in a loss of accuracy if any non-deterministic (within the scope of modelling) component is included in the simulation.

2.4 Emulation

The introduction of operating-system level virtualisation (containerisation) link emulation tools in Linux, FreeBSD and Solaris introduces the possibility to create experimental networks using operating system components. Containers have little overhead and provide a range of isolation methods to assist in the creation of experimental networks. Container-Based Emulators exploit these components to provide the user with a convenient environment to create network configurations. The use of containerisation has some disadvantages as there is a marked loss of fidelity in network traffic throughput and metrics as node density is increased. Pressure on random access memory could result in adverse effects during experiment with high node count. Systems using these techniques are often referred to as emulation systems.

3. Container-Based emulator implementations

Within the context of this paper, a Container-Based Emulator is defined as a purpose made suite of utilities and applications that abstracts the complexity of creating networked containers and enables a user to define and instantiate a set of networked containers, with the ability to define and control configuration values of each deployed component, through a single interface. In this section a set of open-source Container-Based Emulators are reviewed and a brief overview of each CBE is given with regards to history, goals and construction.

3.1 Mininet

MiniNet (Lantz, Heller, and McKeown, 2010) started out as a project to enable large scale OpenFlow (McKeown, Anderson, Balakrishnan, Parulkar, Peterson, Rexford, Shenker, and Turner, 2008) experimentation on commodity hardware. The MiniNet project was then expanded to increase functional realism of network simulations, which resulted in Mininet-HiFi (Handigol, Heller, Jeyakumar, Lantz, and McKeown, 2012; Heller, 2013). The MiniNet GUI exposes a minimal set of components to construct a network topology. The base

components, a host, an openflow switch and controller, a basic switch and basic router is provided. The GUI offers options to provide minimal configuration of each component. Constructing a network topology using the command line tools and configuration files allows the user to exert greater control over network topology and component configuration.

3.2 IMUNES

The Integrated Multiprotocol Network Emulator/Simulator (IMUNES) (Zec and Mikuc, 2004; Salopek, Vasic, Zec, Mikuc, Vasarevic, and Koncar, 2014) is actively developed by the University of Zagreb. Development of the concepts used IMUNES can be traced back to 2002 to a project to virtualise the BSD network stack (Zec, 2002). The goal of IMUNES is to be an alternative to computer network testbeds that can be used on commodity hardware. The IMUNES GUI is developed as an extensive network topology configuration tool. The default component provided include basic networking components such as a hub, switch, router, a PC and a host, as well as a click router and switch. Each component provided can be extensively configured through the user interface.

3.3 Common open research emulator

The Common Open Research Emulator (CORE) (Ahrenholz, 2010) started off as a fork of IMUNES by the United States Naval Research Laboratory (NRL) and Boeing, and is maintained by the Networks and Communication Systems Branch of the NRL. The CORE project extended IMUNES with the ability to execute on Linux, a remote procedure call (RPC) application programming interface (API), a Python library and various user interface (UI) enhancements. Additional goals of the CORE project are to allow wireless network experiments through the Extendable Mobile Ad-hoc Network Emulator (EMANE) (Ahrenholz, Goff, and Adamson, 2011), and the ability to distribute network emulation across multiple hosts. The components provided through the CORE GUI is arranged into two different sets, Layer 2 and Layer 3 components. Layer 2 components include a hub, switch, a wireless LAN emulator and physical Ethernet connections. Layer 3 nodes include a host, PC, a router and a MANER Designated Router (MDR). Layer 3 nodes in CORE can be configured to execute pre-defined service as well as user defined services.

3.4 NetKit and NetKit-NG

NetKit (Pizzonia and Rimondini, 2008; Pizzonia and Rimondini, 2014; Rimondini, 2007) is a project by the Computer Networks Laboratory of the Roma Tre University to enable network experiments to be executed on commodity hardware. NetKit-NG (Iguchi-Cartigny, 2014) is a fork of Netkit, aiming to update the operating system version used. NetKit and NetKit-NG does not provide a GUI, however 3rd party tools such as Visual Netkit Fazio and Minasi, 2009 are available.

3.5 VNX and VNUML

Virtual Networks over LinuX (VNX) (Fernandez, Cordero, Somavilla, Rodriguez, Corchero, Tarrafeta, and Galan, 2011) is a continuation of the Virtual Network User Mode Linux (VNUML) project (Galan, Fernandez, Ruiz, Walid, and Miguel, 2004). VNUML started as an emulation platform to study the address assignment model in IPv6 (Fernandez, De Miguel, and Galan, 2004). The emulation platform used for the study was developed into VNUML to support research project related to computer networks. Development of the VNUML platform was halted in 2009 and has been replaced by VNX. The goals of VNX is to include virtualisation tools to support operating systems other than the host platform in network experiments. It incorporates libvirt and DynaMIPS to achieve these goals. VNX does not have a graphical user interface, however it can produce a graphical map of the current emulation.

3.6 Topology management tool

The Topology Management Tool (ToMaTo) (Schwerdel, Hock, Günther, Reuther, Müller, and Tran-Gia, 2012) is developed by a multi-party group as part of the German-Lab. The goal of ToMaTo is to be a general network experiment environment. Network components in ToMaTo are organised into Devices and Connectors. Devices in ToMaTo emulate networked machines that send and receive packets. Connectors in ToMaTo emulate the links that connect devices and emulate link characteristics.

3.7 Marionnet

The Marionnet project (Loddo and Saiu, 2007; Loddo and Saiu, 2008) was developed by Jean-Vincent Loddo as a teaching aid for his course in networking at the Universit Paris 13. Network components in Marionnet is organised into virtual computers and virtual network devices. The virtual computer components emulate networked machines on the emulated network and virtual network devices emulates hubs, switches, routers and links in the emulated network. A virtual external socket component is provided to link physical Ethernet ports on the host machine to the emulated network. Marionnet provides a desktop application user interface that allows the end user to configure each component in detail.

3.8 Cloonix

The Cloonix network emulation tool (Rehunathan, Bhatti, Perrier, and Hui, 2011; Perrier, 2015) was created by V. Perrier as a tool to assist in the automated creation of simulated network. In contrast to CBE's, Cloonix exclusively uses KVM to virtualise computers, with no option to utilise containerisation techniques. Cloonix has the same objectives as CBE's, a network experimentation platform capable of running on a single host machine, and is thus included. Network components in Cloonix are organised into lan, kvm, tap, c2c and snf objects. The lan object emulates network equipment such as hubs, switches and Ethernet connections, with the kvm component being responsible for creating guest machines in the instantiated topology. The tap, c2c and snf tools provide additional features to connect to physical Ethernet ports on the host, Cloonix to Cloonix distributed emulation and packet sniffing, respectively. The Cloonix user interface provides the user with basic network topology creation tools. Creating a network topology using configuration allows fine grained control over each component in the network topology and allows scripts to be executed on each component during start-up.

4. Container-Based emulator comparison

Each of the CBEs that forms part of this study will be compared in detail using two different comparison methods. The first part will take a look at the architectural choices made to construct the CBE. The second part will compare the specific technologies used to implement the CBE framework.

4.1 Architecture

The architectural choices made during the implementation of each CBE is analysed and compared to better understand the current state of CBEs as frameworks for network experimentation. Each CBE is analysed to assess choices regarding the human-machine interface, how it exposes backend functionality, the design of the backend and the choice of virtualisation technologies. An additional comparison that is included is the capability of a CBE to distribute an emulation across multiple host machines. In Table 1, a preliminary model is shown that will be used to analyse each CBE.

User Interface Each CBE is compared on the interface that it provides to the end user. CBEs that expose only command line (CLI) based interfaces are more difficult to use for beginners but could provide more control for experienced users, whereas CBEs that expose graphical user interfaces (GUI) lower the entry barrier, making them usable for a wider audience. Some CBE implementations expose both GUI and CLI capabilities.

Remote Control The architecture of each CBE is analysed to assess the capabilities of the backend component. The remote control level determines if the CBE framework exposes a remote procedure call (RPC) API, resulting in a loosely coupled architecture. This type of architecture will most commonly expose a backend API library that can be integrated into a CLI and enable distributed emulation. The alternative is a monolithic design, where the exposed user interface (either GUI or CLI) directly controls instantiation of the emulation network.

Virtualisation The comparison of CBE virtualisation techniques assesses the type of technology used to instantiate nodes within the network topology. Each CBE implementation can either use only containerisation or use containerisation and virtualisation. An outlier is Cloonix, which uses full virtualisation, it is included as it is built for the same purpose as CBEs, although it uses only virtualisation technology (KVM).

Distributed Emulation CBEs are compared with respect to the capability of the emulation to be distributed across multiple host machines. Distributed emulation enables the emulation capacity to be expanded horizontally. Distributed emulation capability of a CBE is dependent on the remote control and backend implementation.

Through an analysis of the literature and source code of the CBEs, it was determined that the software packages reviewed shared common architectural designs. In Table 1 the results of the review are shown.

Table 1: Container-Based emulator architecture comparison

Implementation	User Interface	Remote Control	Virtualisation Library	Virtualisation Technology	Distributed
CORE	GUI, CLI	Binary RPC	API Library	Container and Virtualisation	Yes
IMUNES	GUI	None	Monolithic	Container	None
Mininet	GUI, CLI	None	API Library	Container	None
VNX	CLI	3rd Party	API Library	Container and Virtualisation	3rd Party
ToMaTo	GUI	XML-RPC	API Library	Container and Virtualisation	Yes
Marionnet	GUI, CLI	None	Monolithic	Container	Multi-instancing
NetKit	GUI, CLI	None	API Library	Container	Multi-instancing
Cloonix	GUI, CLI	Remore Access	API Library	Virtualisation	Yes

From the tabled results, each of the sections of the architectural model (Table 1) can be subdivided into common approaches used. The user interface component for the all but one of the CBEs is based on a graphical user interface (GUI). VNX does not by default have a user interface, but can create a graphical map of the emulated network through ImageMagick. Half of the reviewed CBEs integrate remote control functionality to allow for distributed emulation. Marrionnet and NetKit achieve distributed emulation through multi-instancing. Only two of the reviewed CBEs (IMUNES and Marionnet) are based on a monolithic design with all other CBEs opting for a component based architecture. CORE, VNX and ToMaTo utilise the strengths of both virtualisation and containerisation to allow for more control of nodes within the experiment if required.

The comparison model of Table 1 is expanded in Table 1 to differentiate between single instance and distributed CBEs.

Table 2: Container-Based emulator architecture model

	CBE Feature	Distributed	Single Instance
System Abstraction	User Interface	Graphical User Interface, Command Line Interface	
	Remote Control	Remote Control Daemon	-
	Virtualisation Library	Application Programming Interface	Monolithic Application, Application Programming Interface
	Virtualisation Technology	Node Emulation, Network Emulation, Link Emulation	

CBEs capable of distributed emulation favours and API model for interacting with containerisation and virtualisation components of the host operating system, though the most significant difference is that single instance CBEs lack remote control capability.

4.2 Implementation

In this section, the technologies that are used to implement the main features of a CBE are enumerated. A CBE has to address a minimum of two aspects, nodes and topology, the base requirements for a computer network. A third aspect of a computer network, link metrics, is addressed by some CBEs. Built in capability to generate background traffic in the emulated network is not addressed in this paper.

Node Emulation The first requirement that a CBE needs to address is that of virtualised nodes within the emulated network topology. Each CBE is analysed with respect to the different technologies that it can use to instantiate nodes. The technologies used to instantiated nodes can range from existing well known systems such as UML, to custom containerisation implementations to address needs specific to the CBE.

Network Emulation The second requirement that a CBE needs to address is the creation of a network topology that links the instantiated nodes. Each CBE is again analysed to determine the technologies used to emulate a

network topology. The network topology instantiated requires a virtual network interface card (NIC) mounted in the emulated node, and a method to connect these NICs to each other, or a virtual switch.

Link Emulation The third component of implementation comparison is link emulation. Link emulation addresses the need to control the characteristics of network traffic flowing in the instantiated topology. The ability to control link metrics such as throughput and packet loss increases the realism (fidelity) of the emulated network, allowing replication of real world conditions for network experiments.

Table 3 lists the version of the CBE that was reviewed, it's host operating system and the different technologies used to implement the CBE.

Table 3: Container-Based emulator technology comparison

Implementation	Version	Operating System	Node Emulation	Network Emulation	Link Emulation
CORE	4.8	Linux	namespaces, LXC, xen	brctl	ebtables
CORE	4.8	FreeBSD	jails	netgraph, vnet	netgraph pipe
IMUNES	2.1.0	FreeBSD	jails	netgraph, vnet	netgraph pipe
IMUNES	2.1.0	Linux	Docker	ovs	
Mininet	2.2.1	Linux	cgroup, netns	ovs, ivs, openflow, brctl	tc, netem
VNX	2.0	Linux	dynamips, libvirt, lxc, uml, vbox, netns	uml switch, ovs, brctl	tc
ToMaTo	3.5.3	Linux	openvz, kvm	tinc vpn, brctl	ipfw (dummynet)
Marionnet	0.90.6 (457)	Linux	uml	vde switch, brctl, tunctl	vde plug
NetKit	2.8	Linux	uml	uml switch	
Cloonix	26.02	Linux	kvm	uml cloonix switch, cloonix mulan	t2t (currently deprecated)

The FreeBSD capable CBEs both use jails and the netgraph system to create emulated networks. Linux is the most popular host operating system for CBEs as all CBEs run on Linux.

UML (User Mode Linux) and KVM (Kernel-based Virtual Machine) is a popular choice as a virtualisation backend with 5 of the 8 CBEs using these technologies. Recent container management systems such as LXC and Docker is by CORE and IMUNES respectively. The only CBE not relying on a 3rd party application to manage containerisation is Mininet. The design goals of Mininet prompted the developers to create a custom container management system to exercise better control over resource usage.

Network emulation on Linux is largely based on Linux bridges (brctl). All CBEs but CORE include support for 3rd party network emulation applications, in particular Open vSwitch. Cloonix is the only CBE to implement custom network emulation technologies. Link emulation is not supported in all CBEs and there is no clear favourite technology for implementing this feature.

5. Conclusion

Available network experimentation platforms give the end user a choice of fidelity level that is most suited to experiments that are to be done. CBEs as an alternative network experimentation platform presents a middle ground in terms of node density and fidelity. CBEs have the ability to have hundreds of nodes in an experimental network while still having access to an operating system kernel capable of executing real world applications. This allows for experimentation that requires interaction with real world applications at a large scale. The open source CBEs reviewed vary in architecture and implementation. These variations in architecture and implementation specifics of the different CBEs reviewed allows the end user to select the most appropriate system based on his or her requirements. For education and training environments, the availability of a graphical user interface supersedes the ability to programmatically control the experimental network. In contrast, experimentation with large scale networks that span multiple computers will benefit from the ability to exercise remote programmatic control over the experimental network. CBEs present a viable, low cost alternative for network administration, education, and security specialists. The specific requirements of an experimental setup will lead the end user to select a CBE that can function within constraints of the environment that the experiment will be executed.

References

Ahrenholz, Jeff (2010). “Comparison of core network emulation platforms”. In: *Military Communications Conference, 2010-MILCOM 2010*. IEEE, pp. 166–171.

Ahrenholz, Jeff, Tom Goff, and Brian Adamson (2011). “Integration of the core and emane network emulators”. In: *Military Communications Conference, 2011-MILCOM 2011*. IEEE, pp. 1870–1875.

Bhatia, Sapan et al. (2008). “Trellis: A platform for building flexible, fast virtual networks on commodity hardware”. In: *Proceedings of the 2008 ACM CoNEXT Conference*. ACM, p. 72.

Fazio, Alessio Di and Paolo Minasi (2009). authors found at http://list.dia.uniroma3.it/pipermail/netkit.users/2008-July/000368.html. url: https://code.google.com/archive/p/visual-netkit/ (visited on 03/08/2016).

Fern´andez, David et al. (2011). “Distributed virtual scenarios over multi-host Linux environments”. In: *Systems and Virtualization Management (SVM), 2011 5th International DMTF Academic Alliance Workshop on*. IEEE, pp. 1–8.

Fern´andez, David, Tom´as De Miguel, and Ferm´ın Gal´an (2004). “Study and emulation of IPv6 Internetexchange-based addressing models”. In: *Communications Magazine, IEEE* 42.1, pp. 105–112.

Galan, F. et al. (2004). “Use of virtualization tools in computer network laboratories”. In: *Information Technology Based Higher Education and Training, 2004. ITHET 2004. Proceedings of the Fifth International Conference on*, pp. 209–214. doi: 10.1109/ITHET.2004.1358165.

Handigol, Nikhil et al. (2012). “Reproducible network experiments using container-based emulation”. In: *Proceedings of the 8th international conference on Emerging networking experiments and technologies*. ACM, pp. 253–264.

Heller, Brandon (2013). “Reproducible Network Research with High-fidelity Emulation”. PhD thesis. Stanford University.

Iguchi-Cartigny, Julien (2014). *NetKit-NG*. url: http://netkit-ng.github.io/ (visited on 03/08/2016).

Lantz, Bob, Brandon Heller, and Nick McKeown (2010). “A network in a laptop: rapid prototyping for software-defined networks”. In: *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. ACM, p. 19.

Linkletter, Brian (2015). *Open-Source Routing and Network Simulation*. url: http://www.brianlinkletter. com/open-source-network-simulators/ (visited on 06/18/2015).

Loddo, Jean-Vincent and Luca Saiu (2008). “Marionnet: a virtual network laboratory and simulation tool”. In: *First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems*.

Loddo, Jean-Vincent and Luca Saiu (2007). “Status report: marionnet or how to implement a virtual network laboratory in six months and be happy”. In: *Proceedings of the 2007 workshop on Workshop on ML*. ACM, pp. 59–70.

McKeown, Nick et al. (2008). “OpenFlow: enabling innovation in campus networks”. In: *ACM SIGCOMM Computer Communication Review* 38.2, pp. 69–74.

Merkel, Dirk (2014). “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux Journal* 2014.239, p. 2.

Perrier, V (2015). *Cloonix*. url: http://clownix.net (visited on 08/02/2015).

Pizzonia, Maurizio and Massimo Rimondini (2008). “Netkit: easy emulation of complex networks on inexpensive hardware”. In: *Proceedings of the 4th International Conference on Testbeds and research infrastructures for the development of networks & communities*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), p. 7.

Pizzonia, Maurizio and Massimo Rimondini (2014). “Netkit: network emulation for education”. In: *Software: Practice and Experience*.

Rehunathan, D et al. (2011). “The study of mobile network protocols with virtual machines”. In: *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 115–124.

Rimondini, Massimo (2007). “Emulation of computer networks with Netkit”. In: *Dipartimento di Informatica e Automazione, Roma Tre University, http://www. netkit. org/, RT-DIA-113-2007*.

Salopek, Denis et al. (2014). “A network testbed for commercial telecommunications product testing”. In: *22nd International Conference on Software, Telecommunications and Computer Networks-SoftCOM 2014*.

Schwerdel, Dennis et al. (2012). “ToMaTo-a network experimentation tool”. In: *Testbeds and Research Infrastructure. Development of Networks and Communities*. Springer, pp. 1–10.

Soltesz, Stephen et al. (2007). “Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors”. In: *ACM SIGOPS Operating Systems Review*. Vol. 41. 3. ACM, pp. 275– 287.

Zec, Marko (2002). “BSD Network stack virtualization”. In: *BSDCon Europe, Amsterdam, Nov* 2.

Zec, Marko and Miljenko Mikuc (2004). “Operating system support for integrated network emulation in imunes”. In: *1st Workshop on Operating System and Architectural Support for the on demand IT InfraStructure (OASIS)*, pp. 3–12.

High Performance Intrusion Detection and Prevention Systems: A Survey

Sasanka Potluri and Christian Diedrich
Otto-von-Guericke University Magdeburg, Germany
sasanka.potluri@ovgu.de
christian.diedrich@ovgu.de

Abstract: There is an enormous growth of industrial applications using internet communication. Secure network is a prime objective for the survival of any organization. Network monitoring and defence systems have become an integral part of network security for identifying and preventing potential attacks. Intrusion Detection and Prevention Systems (IDPS) are network based defence systems which combines Intrusion Detection System (IDS) and a firewall. In contrast to IDS, IDPS is a proactive technique which provides both quick reactions to potential threats and attacks in a network as well as preventing the attacks from entering the network. Current generation IDPS have their limitations on their performance and effectiveness. Some studies have proven that the modern IDPS have difficulties in dealing with high-speed network traffic. Meeting the current network requirements there exist several research approaches to find an efficient IDPS. Nevertheless, serious security and privacy breaches still occur every day, creating an absolute necessity to provide secure and safe information security systems. This survey provides an up-to-date comprehensive review on state of the art of IDPS based on different accelerating techniques, different detection algorithms, types of hardware and optimizing algorithms to match the demand requirements of high speed network. A detailed overview on high performance IDS and IDPS along with pros and cons of individual techniques will be given. This paper also highlights and discusses the requirement for developing a new IDPS to detect the known and unknown threats.

Keywords: intrusion detection system, intrusion prevention systems, high performance computing, network security

1. Introduction

Today's businesses extremely rely on corporate IT networks and their connections with the global Internet. The number of individuals using the internet increase from 1 billion in 2005 to over 2.7 billion in 2013 (International Telecommunications Union (ITU) 2013). Cisco (2013) estimates that global internet traffic has increased from 2000 Gbps in 2007 to 12,000 Gbps in 2012 and forecasts that this will increase to 35,000 Gbps by 2017, equivalent to 1 zettabyte per year. Global IP traffic has increase more than fivefold in the past 5 years, and will increase nearly threefold over the next 5 years. Annual global IP traffic will surpass the zettabyte (1000 Exabyte's) threshold in 2016, and the two zettabyte threshold in 2019. The number of devices connected to IP networks will be three times as high as the global population in 2019 and the broad band speeds will double by 2019 and will reach 43 Mbps, up from 20Mbps in 2014 (Cisco 2015). The above statistics show the evolution of internet in our daily life. With recent advancements in the internet speed many industrial infrastructures are relying on internet based device communication rather than the cooperate infrastructure. This gains the advantage of easy installation and establishment on the other side security threats are raising which can be considered as an economical threat to the industry.

1.1 Background

Approximately 7.6 million new unique pieces of malware were detected by the AV-Test Institute for the month of June 2013 (AV Test 2013). In other words, a new malware was created every 0.35 seconds. However, the convenience of global connectivity comes at a cost – the vulnerability of the network and systems to the malicious actions of cyber criminals. According to a survey by The Global State of Information Security Survey 2016 (PWC 2016) 91% follow a risk-based cybersecurity framework, 69% use cloud based cybersecurity services, 59% leverage Big Data to improve cybersecurity, and 65% collaborate with others to improve cybersecurity. In 2015 38% more security incidents were detected than in 2014. Theft of hard intellectual property increase 56% in 2015. Respondents boosted their information security budgets by 24% in 2015. 48% are actively monitoring/ analysis of security intelligent to safeguard their ecosystems against evolving threats. The survey result is based on responses of more than 10,000 CEOs, CFOs, CIOs, CSOs, VPs and directors of IT and security practices from more than 127 countries.

From another survey (HIMSS 2015) it says that 66% of organizations had experienced a security incident. 87% of respondents indicated that information security had become a critical business priority. 81% of respondents believe more innovative and advanced tools are needed to combat security threats. From (Skybox Security 2013)

46% North American Organizations and 60% of European organizations manage more than 100 rules per firewall – Europeans reported more than twice the number of rule changes per month. 93% of organizations use or plan IPS modules for their Next Generation Firewall (NGFW); 62% with active management.

1.2 Why intrusion detection and prevention systems?

With the rapid application of the network technology in industries, the problems with network security also appears to be serious. The traditional Firewall technologies can't provide the enough security against the novel attacks and intrusions (Patel et al. 2010). Firewalls are designed to restrict unauthorized data transmission to and from the network. Although this can safeguard the network from certain intrusions, it also has several vulnerabilities. Firewall does not detect attacks dynamically. A firewall is just a fence around a network designed to block certain types of communications routed or passing through certain ports. It is not designed to discover someone bypassing the firewall or digging a tunnel under the fence. Moreover, an IDS has the capability to detect a broader range of attacks from within the network, whereas the firewall cannot. An IDS has the capability to detect insider attacks, look at misconfigured firewalls and most important, capture information on failed attempts. Apart from firewalls, the IDS systems are faced with compromise between false alarms and false positives (Brown et al. 2002). This makes the researchers look into IDPS.

2. Goals of intrusion detection and intrusion prevention

Attacks on industrial systems are performed by threat actors with varying sophistication and goals. While it is not possible to list all the potential attacks, it is good to know that the list is always growing. Some types of attacks on communication network are vulnerabilities (Martin 2001), SYN flooding, Distributed Denial-of-Service (DDOs), surfing (Anderson 2008) and the list goes on. Intrusion detection refers to the detection of malicious activity (attacks, break-ins, penetrations and other forms of computer abuse) in a computer related systems or in the communication networks. An intrusion can be sometimes identified as a completely different behaviour from the normal and sometimes hard to identify it from normal behaviour. These malicious activities or intrusions are very important in network security perspective. Due to these complexities there doesn't exists a unique technic that can identify all types of attacks or intrusions.

As IDS deals with only detection, it is considered as a passive mechanism only. In order to prevent attacks, we need systems that can detect attacks online and prevent malicious data to enter the network. Hence IDS is improvised as IPS (Intrusion Prevention System) in some literature this is also termed as IDPS (Intrusion Detection and Prevention Systems). An overview on intrusion detection and intrusion prevention systems is given below.

2.1 Characteristics of intrusion detection system

An IDS is a software or a hardware that helps to protect from and ward off attacks and penetration attempts to the network. The key difference from an IDS to a simple firewall or adaptive proxy firewall is that the firewalls can block connections. IDS is a combination of several individual intrusion detection techniques available (signature analysis, traffic monitoring and anomaly detection). IDS checks the network behaviour and finds the nodes that are not working normally. It is an additional unit installed at the clients or server or both (Farooqi et al. 2009).

2.1.1 Types of intrusion detection system

NIDS – Network based Intrusion Detection System.

NIDS are placed at an intentional point or points with in the network to monitor the traffic going in and out of all devices in the network (Ajay et al. 2014). NIDS are mostly passive devices that monitor the on-going network activity without adding significant overhead or interfering with the network operations. They are easy to secure against attack and may even be undetectable to attackers; they also require little effort to install and use on existing networks. Ideally it would scan all inbound and outbound traffic; however, doing so might create a bottleneck that would impair the overall speed of the network.

HIDS – Host based Intrusion Detection System

HIDS runs on individual hosts or devices on the network. A HIDS monitors the incoming and outgoing packets from the device only and will alert the user or administrator of suspicious activity detected (Ajay et al. 2014). The suspicious activities are based on the type detection technique employed. For example, audit analyse technique is able to identify the activities related to operating-system-level intrusion and application-level intrusions.

Signature based Intrusion Detection System

A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats (Ajay et al. 2014). This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS: During that lag time your IDS would be unable to detect the new threat.

Anomaly Intrusion Detection System

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different than the baseline (Ajay et al. 2014).

2.2 Characteristics of intrusion prevention system

IPS is an advanced form and is a combination of IDS, firewalls etc. The purpose of an IPS is not only to detect an attack that is trying to interrupt, but also to stop it by responding automatically such as logging off the user, shutting down the system, stopping the process and disabling the connection etc. IPS today are extremely effective and scalable because they perform deep packet inspection to ensure that only legitimate traffic makes it into the network. IPS solutions are considered as an active mechanism to fight against intruders in contrast to IDS which is considered as a passive detection mechanism.

2.2.1 Types of intrusion prevention system

HIPS – Host Based Intrusion Prevention System

It is the merge of both IDS and firewall on a single system. HIPS relates to the processing of data that originated on the computers themselves, such as event and kernel logs. HIPS also monitors which program accesses which resources and might be flagged (Sandhu et al. 2011). It also monitors the state of the system and try to find anomalies based on filter technics. HIPS maintains a database of system objects and also stores the system's normal and abnormal behaviour. The database contains important information about system files, behaviour and objects such as attributes, modification time, size etc. If any suspicious or anomaly behaviour occurs, then it generates an alarm and takes some appropriate response against detected threat or attack.

NIPS – Network Based Intrusion Prevention System

When the IPS is used to analyse the network packets then it is called as NIPS (Sandhu et al. 2011). It captures the each and every network traffic from the connection as it travels to a host. This can be analysed for a particular signature or for unusual or abnormal behaviours. Several sensors are used to sniff the packets on network which are basically computer systems designed to monitor the network traffic. If any suspicious or anomaly behaviour occurs, then they trigger an alarm and pass the message to the central computer system or administrator and then an automatic response is generated. There are further two types of NIPS. First, promiscuous-mode network intrusion detection, which is a standard technique that sniffs all the packets on a network segment to analyse the behaviour. In this mode only one sensor is placed on each segment in the network. Second, network-node intrusion detection system that sniffs the packets which are bound to a particular destination computer. These are designed to work in a distributed environment (Kozushko 2003).

2.3 Intrusion detection system vs intrusion prevention system

The system overview of IDS and IPS along with key differences is shown in Figure 1.

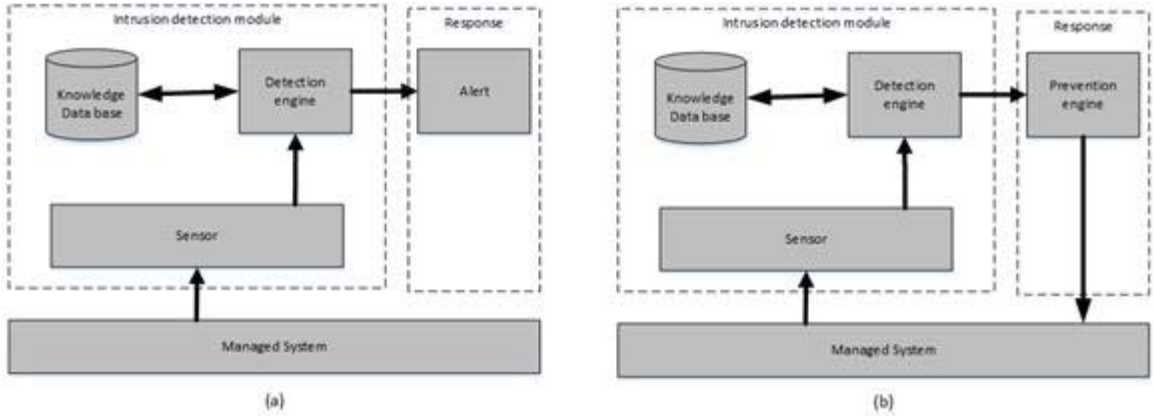


Figure 1: (a) intrusion detection system (b) intrusion prevention system

As shown in the Figure 1, the components present in the detection module of IDS and IPS were the same. The sensors in the detection module is used to monitor and analyze the network traffic (toward or away from the network). The detection engines tries to analyze and identfy the activities happening in overall network or within the individual host. The detector uses three different sources of information: A permanent database (knowledge base or normal beaviours profiles) used to identify malicious or intruding activite, Audits (information about activitites going on within the system) and systems current configuration (used in the later stage to reinstate the system) (Mohammad et al. 2010).

The key difference exits in their response mechanism. The main function of an IDS is to warn or generate an alert when some suspicious activity is identifies while IPS is designed and developed for more active protection to improve upon the IDS and other traditional security solutions, which can react in real time to block or prevent those activites whithout any external intervention. The response of the IPS on detected threats can be in severral ways, for example:

- It can reconfigure other security control in systems such as firewall or router to block the future attacks
- It can remove malicious content of an attack in network traffic to filter out the threatening packets; or
- It can (re-)configure other security and privacy controls in browser settings to prevent future attacks.

Different IDS/IPS takes different approaches to identify the attacks. There exist some common issues that plague the range of detection strategies. Some challenging issues in developing the IDS/IPS include rule set for attack identification (Brown et al. 2002), disparities in detecting attacks, training behavioural models, attacks against the IDS, passive or active protection (Svein et al. 2007), gigabit network are some of them.

3. High performance for intrusion detection and prevention

As explained above the main task in the IDS/IPS technology is to look at the network packets for malicious activities. These systems consist of a set of rules that are mainly compared against the network packets to identify matched patterns. For example, a signature based IDS is configured with thousands of rules that detect malicious attacks and codes. Usual rules consist of a filter specification based on packet header fields, a string that must be contained in the packet payload, the approximate or absolute location were that string should be present, and to take an associated action if all the conditions of the rule are met. Signature matching is a highly computationally intensive process, accounting for about 75% of the total CPU processing time of modern NIDS (Vasiliadis et al. 2008). While working with IDS the most tedious task is sifting through the large volume of data generated. IPS solution stay inline on a network just as a firewall or router does. As a consequence of sitting in-line, IPS solutions need to match with high network traffic or gigabit-level requirements. While today's data IPS technology is a significant step forward, the debate continues. Along with accuracy requirements of IPS, it also has to minimize their latency. It is considered as one of the major challenge since inline IPS available today introduce latency as every packet needs to be inspected. The obvious need for increased accuracy to handle the high volume of data in IPS systems is rapidly raising the performance bar for IPS products in the market place today.

As a need to speed up the inspection process, the search for high performance solution is necessary. Some studies (Clark et al. 2005) also shown that the modern NIDS have difficulty in dealing with high speed network traffic. Others (Ptacek & Newsham 1998) have shown how attackers can use this fact to hide their exploits by overloading an NIDS with extraneous information while executing an attack. The high performance can be achieved by modifying existing software or a hardware or by a combination of both of them. Software acceleration is the process of improving the algorithm's efficiency using the multicore capability of CPU's and by multi-threaded programming running on Personal Computers (PC's) or using a computing paradigm which supports parallel processing (E.g. Neural Networks). Hardware acceleration is the process of computing the computer intensive tasks on a special multicore platform to achieve faster computation capabilities. Example for hardware acceleration includes Graphic Processing Units (GPU's) (Vasiliadis et al. 2008), Field Programmable Gate array (FPGA's) (Chandy 2004), Application Specific Integrated Circuits (AISC's) (Lu et al. 2006).

There also exits a hybrid acceleration mechanism where the use of both software computing paradigm and hardware acceleration capabilities or used to accelerate the process (Foschini et al. 2008). There exist several researchers who have increased the performance of the intrusion detection and prevention. In the next section we provide a thorough review of the existing research. We also provide a summary of recent advancements in the technology development of the IDS and IPS technology.

4. Literature review

The authors of (Sekar et al. 1999) aim to develop a new approach for NIDS based on concise specification that characterize normal and abnormal network packet sequences. The key feature of this implementation is a domain specific language for capturing patterns on normal and/or abnormal network packet sequences. This language supports concise and easy-to write attack patterns which in turn increase confidence in attack specifications and reduces the development and debugging times needed for defending against new attacks.

(Lu et al. 2006) provides a new memory efficient multiple-character-approaching architecture suited for AISC implementations called Transition-Distributed Parallel Deterministic Finite Automata (TDP-DFA). They introduced parallel DFA's with overlapping input windows to achieve the goal of processing multiple characters in each clock cycle. Switching Fabric based structure and Bloom filter based classifier for sharing most of the transition rules among them significantly reduce the overall storage cost.

Rather than optimizing the pattern matching algorithms, parallelizing the signature matching process performs the network detection process faster. This was discussed in (Foschini et al. 2008) where they provided a stateful signature matching that has been implemented only using the off-the shelf components. Efficient pattern matching is considered as a key issue in NIDS. Hence network-based scheme practically requires an efficient algorithm suitable for hardware implementations. In (Yu et al. 2004) they developed a Ternary Content Addressable Memory (TCAM) based multiple pattern matching scheme. This scheme can handle complex patterns such as arbitrarily long patterns, correlated patterns and patterns with negation.

To cope with higher traffic throughput and increasing link speed some researchers used hardware accelerators to perform the NIDS. (Das et al. 2008) uses FPGA based architecture for anomaly detection in network transmission. They developed a Feature Extraction Module (FEM) which aims at summarizing network information to be used at later stage of NIDS. They used Principal Component Analysis (PCA) as an outlier detection method for NIDS. Where as in (Vasiliadis et al. 2008) used the underutilized computational power of modern GPU's to offload the costly pattern matching operations form the CPU and thus increase the overall processing throughput. (Lunteren 2006) designed a novel scheme for pattern-matching, called BFPM (BFSM based patter matching), that exploits a hardware based programmable state machine technology to achieve deterministic processing rates that are independent of input and pattern characteristics in the order of 10Gbps for FPGA and at least 20Gbps for AISC implementations. BFSM stands for Balanced Routing Table (BaRT) based Finite State Machine. BFPM combines deterministic pattern matching performance, fast dynamic updates and several other features that are important to NIDS.

Soft-computing and machine learning techniques are rigorously used to build autonomous IDS. A survey on machine learning techniques for *IDS* is given in (Singh & Nene 2013). Neural networks one of the most commonly used machine learning technique is used intensively in developing IDS. (Bastke 2009) uses probabilistic neural networks which can be massively parallelized were used to implement the IDS on GPU's. The combination of

statistical network data, self-learning algorithms and computation power of GPU's makes the system to adapt independently to different environments. (OuYang 2011) uses GPU based Reduction Support Vector Machine algorithm (GPU-RSVM) which include three parts: SVM training data reductions, SVM based intrusion detection classification model training and intrusion detection testing.

As software based NIDS are too compute intensive and cannot meet the bandwidth requirements of a modern network (Chandy 2004) proposes a FPGA based keyword match processor that can serve as the core of a hardware based NIDS. The keyword match processor's key feature is a cellular processor architecture that allows Content Addressable Memory (CAM) to process variable sized keys. Kargus, a highly scalable software based IDS (Jamshed et al. 2012) takes the advantage of all available system resources. It improves the performance by realizing two key principles: batching and parallelism. It exploits high parallelism in modern computing hardware by efficiently balancing the load of flow across multiple CPU cores and by employing a large array of GPU processing cores. Kargus achieves 33Gbps for normal traffic and 9 to 10Gbps even when all traffic is malicious.

Recent advancements in IPS also made them to look in the direction of high performance IPS. (Artan et al. 2007) developed network IDPS (NIDPS) search for certain malicious content in network traffic (i.e. signatures). They presented a 10 Gbps hardware NIDPS and achieved high speed detection using single FPGA without any external memory. Regular expression (RE) matching is a core component of deep packet inspection in modern networking and security devices. (Meiners et al. 2010) proposes the first hardware-based RE matching approach that uses TCAM's which are off the shelf chips and have been widely deployed in modern networking devices for packet classification. Three techniques: transition sharing, table consolidation and variable striding were used to reduce TCAM space and improve RE matching speed. Centralized NIDS have various limitations on their performance and effectiveness. To overcome the drawbacks of centralized NIDS (Clark et al. 2005) argues that intrusion detection analysis should be distributed to Network Node IDS (NNIDS). An NNIDS can unambiguously inspect traffic to and from the host and when implemented on the network interface hardware, can function independently of the network interface hardware, can function independently of the host operating system to provide better protection with less overhead than software implementations.

To overcome the trade-off between the accuracy of detection and algorithmic efficiency (Weinsberg et al. 2006) proposes a novel pattern matching algorithms named as Rotating TCAM (RTCAM). This algorithm is capable of matching multiple patterns in a single operation. RTCAM enables the NIPS appliance to operate at an aggregate rate of several gigabits per second. Another high performance NIPS that combines the use of software-based network intrusion prevention sensors and a network processor board was discussed in (Xinidis et al. 2005). The network processor acts as a customized load balancing splitter that cooperates with a set of modified content-based network intrusion detection sensors in processing network traffic. They showed that the components of such a system, if co-designed can achieve high performance, while minimizing redundant processing and communication. A cheaper and readily scalable to future high speeds and retain the unparalleled flexibility of IPS system was mentioned in (Weaver et al. 2007). The Shunting architecture uses a simple in-line hardware elements that maintains several large state tables indexed by packet header field including IP/TCP flags, source and destination IP addresses and connection tuples. They aimed at implementing an FPGA-based realization of Shunting.

Table 1: Critical analysis of high performance intrusion detection and prevention systems

Literature	Accelerator type	Device	Detection	Prevention	Technique	Pros	Cons
(Sekar et al. 1999)	Software	PC	Yes	No	Pattern Matching	Easy to write attack patterns, decrease pattern matching time	No mechanism of protection, Performance up to 500 Mbps only
(Lu et al. 2006)	Software/ Hardware	PC/ASIC	Yes	No	Pattern Matching, TDP-DFA	Minimizing the transition rules and storage costs	Cannot detect anomaly behavior intrusions
(Foschini et al. 2008)	Software/ Hardware	Parallel architecture	Yes	No	Parallel matching	Parallel and stateful	No mechanism of

Literature	Accelerator type	Device	Detection	Prevention	Technique	Pros	Cons
		/PC			algorithm, Snort	intrusion detection for high speed networks	protection, Suitable to know attacks
(Yu et al. 2004)	Hardware	TCAM	Yes	No	Pattern Matching	Can operate at 2Gbps rate, is efficient for long patterns	No mechanism of protection, Suitable to know attacks
(Das et al. 2008)	Hardware/ Software	FPGA/ PCA	Yes	No	PCA, FEM	Improved network throughput, pipelining and hardware parallelism	Dimensionality reduction may not fit to detect all outliers
(Vasiliadis et al. 2008)	Hardware	GPU	Yes	No	Aho-Corasick string matching algorithm, Snort	Achieve faster pattern matching of 2.3 Gbps	No mechanism of protection, Packet drop increases with increase in data speed
(Lunteren 2006)	Software/ Hardware	PC/FPGA/ASIC	Yes	No	Pattern Matching, Dynamic updates	High deterministic performance, high storage efficiency, high processing rates	State flows for a complex scenario are hard to handle
(Bastke 2009)	Hardware	GPU	Yes	No	Machine Learning	Adaptability to new environments	Lack of proper benchmarking and evaluation
(OuYang 2011)	Hardware	GPU	Yes	No	Machine Learning	Cost time of reduction process decreases greatly	Prediction accuracy hasn't obviously declined
(Chandy 2004)	Hardware	FPGA	Yes	No	Key word matching, CAM	Can meet the demands of future up to 10Gbps	Unable to detect anomaly behaviour
(Jamshed et al. 2012)	Software	CPU/GPU	Yes	No	Batching, parallelism	Achieve speed up to 33Gbps	CPU to GPU offloading needs to carefully analyzed
(Artan et al. 2007)	Hardware	FPGA	Yes	Yes	Signature based Detection	Achieve faster signature mapping up to 10 Gbps	Detecting signature over multiple packets is missing
(Meiners et al. 2010)	Hardware	TCAM	Yes	Yes	RE matching	A throughput of up to 18.6 Gbps is possible	Concentrate mainly on space saving
(Clark et al. 2005)	Hardware	FPGA	Yes	Yes	Distributed and	Achieved high speed	Anomaly intrusion

Literature	Accelerator type	Device	Detection	Prevention	Technique	Pros	Cons
					collaborative NNIDS based on Snort	intrusion detection and prevention	detection is not considered
(Weinsberg et al. 2006)	Hardware	TCAM	Yes	Yes	Pattern Matching	Snort compatibility, an average line speed of 12.35 Gbps	Memory issues in implementing RTCAM
(Xinidis et al. 2005)	Hardware/ Software	Network Processor/ PC	Yes	Yes	Pattern Matching	Improved performance in packet load balancing	Performance is less than 1 Mbps and several software components
(Weaver et al. 2007)	Hardware	FPGA/ NetFPGA	Yes	Yes	Shunting, RNET	Track connections and addresses of interest	Types of intrusion detection is not discussed

5. Conclusion and future work

In this paper, we introduced the need of intrusion detection and prevention system followed by the requirements of high performance in IDPS. Our survey includes the challenges, existing and ongoing research related to intrusion detection and prevention. Different techniques discussed in this paper support the security of an organization or a computer network against threats or attacks. The attackers, on the other side realising the new ways of breaching the security measures. The best way is to develop a strongest model or mechanism which provides the best protection against known and unknown threats and ensure system is secure. Different techniques used by IDPS like pattern matching, string matching, machine learning techniques and hardware computation power were used to provide a great and strongest security against numerous attacks. Finally, we believe that this area of research is still active and several works need to be performed on the different sides of the implementation (hardware and software) that meet the increasing demand. In future, we will propose a solution to further secure organisation network without compromising the network's performance.

Acknowledgements

This work is being supported by the project isSecure funded by Federal Ministry for Economic Affairs and Energy (BMWi) and Federation of Industrial Research Associations "Otto von Guericke" (AIF), a collaborative project between Otto-von-Guericke University Magdeburg (Institute for Automation Engineering) and ifak systems GmbH, Magdeburg.

References

Ajay kaurav, S.Sibi Chakkaravarthy, R.Patil, Pravin, M.V.K., (2014). Intrusion Detection system: A Review of the state of the art. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 16(1), pp.108–112.

Anderson, R., (2008). Nework Attack and Defense. In *Security Engineering*. Wiley.

Artan, N.S. et al., (2007). A 10-Gbps High-Speed Single-Chip Network Intrusion Detection and Prevention System. *IEEE GLOBECOM 2007-2007 IEEE Global Telecommunications Conference*, pp.343–348.

AV Test, (2013). The best antivirus software for Windows Client. , p.June 2013. Available at: https://www.av-test.org/en/antivirus/home-windows/windows-7/june-2013/ [Accessed February 3, 2016].

Bastke, S., (2009). Combining statistical network data, probabilistic neural networks and the computational power of GPUs for anomaly detection in computer networks. *Workshop Intelligent Security (SecArt 2009)*, (iii), pp.1–6.

Brown, D.J., Suckow, B. & Wang, T., (2002). A Survey of Intrusion Detection Systems. *Department of Computer Science, University of California, San Diego*.

Chandy, J. a., (2004). FPGA Based Network Intrusion Detection using Content Addressable Memories. *12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines*, pp.316–317.

Cisco, (2015). *Cisco Visual Networking Index: Forecast and Methodology, 2014-2019*,

Clark, C. et al., (2005). A hardware platform for network intrusion detection and prevention. In *Proceedings of the 3rd Workshop on Network Processors and Applications*.

Das, A. et al., (2008). An FPGA-based network intrusion detection architecture. *IEEE Transactions on Information Forensics and Security*, 3(1), pp.118–132.

Farooqi, A.H. et al., (2009). Intrusion Detection Systems for Wireless Sensor Networks: A Survey. , 56, pp.234–241.

Foschini, L. et al., (2008). A parallel architecture for stateful, high-speed intrusion detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5352 LNCS, pp.203–220.

HIMSS, (2015). *HIMSS Cybersecurity Survey 2015*,

International Telecommunications Union (ITU), (2013). ITU (2013). ICT indicators for developed and developing countries and the world (totals and penetration rates). Available at: www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/ITU_Key_2005-2013_ICT_data.xls [Accessed February 1, 2016].

Jamshed, M. et al., (2012). Kargus: A Highly-scalable Software-based Intrusion Detection System. *ACM Conference on Computer and Communications Security*, pp.1–12.

Kozushko, H., (2003). Intrusion detection: Host-based and network-based intrusion detection systems. , 11.

Lu, H. et al., (2006). A memory-efficient parallel string matching architecture for high-speed intrusion detection. *IEEE Journal on Selected Areas in Communications*, 24(10), pp.1793–1803.

Lunteren, J. Van, (2006). High-Performance Pattern-Matching for Intrusion Detection. , 00(c).

Martin, R.A., (2001). Managing vulnerabilities in networked systems. *Computer*, 34(11).

Meiners, C.R. et al., (2010). Fast regular expression matching using small TCAMs for network intrusion detection and prevention systems. *Proceedings of the 19th USENIX conference on Security*, pp.111–126.

Mohammad Naveed, Muhammad and Un Nihar, S. and I.B., (2010). Network intrusion prevention by configuring ACLs on the routers, based on snort IDS alerts. In *6th International Conference on Emerging Technologies (ICET)*. IEEE, pp. 234 – 239.

OuYang, Q., (2011). Theoretical and Mathematical Foundations of Computer Science. *Communications in Computer and Information Science*, 164(January), pp.154–160.

Patel, A., Qassim, Q. & Wills, C., (2010). A survey of intrusion detection and prevention systems. *Information Management & Computer Security*, 18(4), pp.277–290.

Ptacek, T. & Newsham, T., (1998). Insertion, Envasion, and Denial of Service: Eluding network intrusion detection.

PWC, (2016). *Turnaround and transformation in cybersecurity*, Available at: www.pwc.com/gsiss.

Sandhu, U.A. et al., (2011). A Survey of Intrusion Detection & Prevention Techniques. *2011 International Conference on Information Communication and Management*, 16, pp.66–71.

Sekar, R. et al., (1999). A high-performance network intrusion detection system. *Proceedings of the 6th ACM conference on Computer and communications security - CCS '99*, pp.8–17.

Singh, J. & Nene, M.J., (2013). A Survey on Machine Learning Techniques for Intrusion Detection Systems. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(11), pp.4349–4355.

Skybox Security, (2013). Skybox Security Survey:Next-Generation Firewall Management. Available at: http://www.informationweek.com/whitepaper/Security/Security-Administration/rule-driven-profiling-a-next-generation-approach-wp1357662164/116533?gset=yes& [Accessed February 1, 2016].

Svein Haslum, Kjetil and Abraham, A. and K., (2007). Dips: A framework for distributed intrusion prediction and prevention using hidden markov models and online fuzzy risk assessment. In *Third International Symposium on Information Assurance and Security*. IEEE, pp. 183 – 190.

Vasiliadis, G. et al., (2008). Gnort: High performance network intrusion detection using graphics processors. *Recent Advances in Intrusion Detection*, pp.116–134.

Weaver, N., Paxson, V. & Gonzalez, J.M., (2007). The shunt: an FPGA-based accelerator for network intrusion prevention. *Proceedings of the 2007 ACM/SIGDA 15th international symposium on Field programmable gate arrays*, pp.199–206. A.

Weinsberg, Y. et al., (2006). High performance string matching algorithm for a network intrusion prevention system (NIPS). *2006 Workshop on High Performance Switching and Routing*, p.7 pp.

Xinidis, K., Anagnostakis, K. & Markatos, E., (2005). Design and implementation of a high-performance network intrusion prevention system. *Security and privacy in the age of ubiquitous computing*, pp.359–374.

Yu, F., Katz, R.H. & Lakshman, T. V., (2004). Gigabit rate packet pattern-matching using TCAM. *Proceedings - International Conference on Network Protocols, ICNP*, pp.174–183.

Cultural Comparison Between and Attackers and Victims

Char Sample and Mardi John
University of Warwick, UK and MITRE, USA
charsample50@gmail.com
mkjohn@mitre.org

Abstract: The linkage between those who deface government owned websites and culture was first established by Sample (2013); however, Karamanian, Sample and Kolenko (2016) added to the body of knowledge when they examined target (or victim) sites for cultural commonalities. The cultural markers for victim sites, while weaker than the attacking counterparts, were present. Some commonalities as well as differences exist between attackers and targets. This study delves deeper into both the commonalities and differences to offer a more comprehensive profile of both groups.

Keywords: cyber actors, behaviours, Hofstede, cultural dimensions, automatic thought

1. Introduction

Beidleman's (2009) observation of the emergence of cyberspace as a setting for war domain continues to hold true. Hayden's (2011) observation about the newness of the domain and the difficulties in transferring knowledge from physical domains to the cyber domain still creates problems for defenders. These problems may be exacerbated when defenders fail to understand the cultural perspective of the attacker. Hayden (2011) observed the inability to accurately attribute hostile acts is unique to the cyber domain. This inability to accurately attribute attacks with high levels of confidence suggests the need for additional attribution and possibly greater threat actor intelligence information in cyberspace.

The ability to model cyber actors' behaviours in cyber space could possibly provide a missing element in the complex attribution puzzle, while allowing defenders the opportunity to simulate the most effective responses. Modelling behaviours will require a framework that incorporates technological, psychological and anthropologic insights. A well-known and widely used anthropologic model that is frequently used in various disciplines is Hofstede's cultural framework (Hofstede, Hofstede, & Minkov, 2010). Originally introduced in the late 1970's, Hofstede's framework remains relevant and continues to be updated, most recently 2013.

In 2013 Sample, used Hofstede's framework, to successfully linked nationalistic, patriotic themed website defacements with the Hofstede defined cultural characteristics of high power distance and collectivism. The 2013 study led to another study where Sample & Karamanian (2015) examined cyber behaviours associated with the domain name system. Liao, Pan & Zhou (2009) linked online preferences and cyber communities to national culture.More recently Sample (2015) examined a larger set of defacements in the context of physical events, resulting in some characterizations of attackers and targeted victims. However, this study will provide a deeper analysis and a more detailed explanation of the findings through a comparison between attackers and targeted victims.

2. Literature review

Gonzales-Vallejo, Lassiter, Bellezza, and Lindberg (2008) noted the role of unconscious thoughts and processing on cognition and behaviour. These unconscious patterns shape the perceptions that alert the person of an event or problem. This finding supports the observation by Bargh and Morsella (2008) that unconscious thought precedes conscious thought. These findings indicate that some behaviors are unconsciously driven. Additionally, these findings support Nisbett's (2010) observation of the influence of priming events on perception.

Bargh and Morsella (2008) also observed the intermingling of conscious and unconscious thought that can be observed when a person types or drives a car. Driving a car presents an interesting example since, part of the process relies on automatic thought, while other times full concentration is required. In both cases the individual began by learning the steps consciously but through repetition both of these procedures are consciously initiated and are unconsciously run (Ibid). The typing example is particularly relevant since keyboards are presently an integral component of the cyber environment. Instance based learning (Gonzalez & Dutt, 2011) provides additional support of the conscience to unconscious thought migration.

While the environment may factor in the shaping of the operator's initial perception, Gonzales-Vallejo et al., (2008) also recognized that the environment in which events occurred was not the only factor required in order to predict the response. A key component in the overall thought process, perception, may be influenced by factors that reside outside of the visible immediate environment. Perception may be sensed before recognized (Ibid). Perception provides the initial realization of an event therefore; the context or environment in which an event is perceived would logically influence the receiver's identification of the event.

The thought environment can be influenced by many factors ranging from the physical environment to individual preferences. However, in this study the authors choose to focus on national culture, the relationship between culture and thought leading to and understanding of culture and behaviours, particularly cyber behaviours. In order to understand the role of culture in cyber behaviours an understanding of the role of culture in behaviours should be examined.

Bargh & Morsella (2008) described the role of culture in a child's world in terms of learning appropriate behaviours. Baumeister and Masicampo (2010) further re-enforced the role of culture in cognition that later becomes unconscious thought. Recognizing that perception is an early stage of cognition, and realizing that culture influences perception (Morris and Peng, 1994; Hofstede et al., 2010; Nisbett, 2010), the earliest warnings to our cyber systems are interpreted by operators who are culturally influenced.

The importance of automatic thought emerges following the analysis of Elmasry, Auter and Peuchaud (2014) of online identity-construction of Facebook users from different cultures. Elmasry et al (2014) observed the selections and postings of information appropriate to country cultures' norms, supporting the unconscious embeddedness of culture in identity selections. Referencing Hofstede's dimensions, the analysis cross-compared Arab and East Asian cultures through Hofstede dimensions' characteristics of low tolerance for change, and norms for masculinity and collectivism versus individualism.

In consequential activities such as selection of infant sleeping arrangement (Shimizu, Park & Greenfield, 2014) to classroom practices (Kaur & Noman, 2015), culturally accepted practices were repeatedly preferred over other available options. Customary norms being selected as defaults did have drawbacks as studied by Borkovich (2012), with the examination of multinational organizations, and the surprising incongruities resulting from mergers and acquisitions. Employees of subsidiary companies of a larger parent company faced culture shocks, counter to the harmonious business environments and the profit motive by employees that were to provide commonalities for garnering an atmosphere of greater efficiencies and cost savings. Employees of subsidiary organizations faced culture shocks in dealing with employees of the larger parent organizations, and ensuing incongruities generated shocks that hampered effectiveness and growth. In analyses, investigators relied on Hofstede's framework, and the underpinning of unconscious cultural contexts that undercut other motivations, as well as environmental influences.

In conflict resolution, Khanaki and Hassanzadeh (2010) and Wei (2001) studied resolution styles for different cultures. Focusing on Hofstede's Individualism-Collectivism dimension, Wei (2001) also observed some influence in conflict resolution amid a cross-cultural environment, but Khanaki and Hassanzadeh (2010) concluded that greater assertiveness in conflicts was rooted in stronger individualism (as per Hofstede's dimension), separating ways by which cultural framework influenced the conflict management styles of individuals in selecting avoidance, accommodation, compromise, competition or collaboration when resolving conflicts. These findings would suggest the possibility that conflict resolution in the cyber domain may exhibit some similarities to conflict resolution in the physical domain.

The findings by Bargh and Morsella (2008) along with Hofstede et al. (2010) and Minkov's observation about the independence of culture (2013) influenced Sample's (2013) study that linked nationalistic, patriotic website defacements (NPWDs) to cultural values as defined by Hofstede. Sample's (2013) use of Hofstede's quantitative framework along with statistical methods used to evaluate cyber behaviours led to several subsequent studies that focused on various cyber actors. The most recent study by Sample (2015) allowed for the characterization of defacers and victims within the context of physical (or kinetic) events to be identified. However, the study while successfully identifying common cultural traits amongst actors did not address, in great detail, explanations for the findings. This study does.

3. Methodology

The website www.zone-h.org provided 10 years of attacker and victim data for use in the Sample (2015) study. The study focused on government owned sites limiting the domains .mil.*cc* and .gov.*cc,* websites, where the country code is represented as *cc.* Hypothesis testing was performed using the Wilcoxon rank sum test to compare attackers to non-attackers. Tests were performed comparing victims to non-victims;

$$W = \sum_{j=1}^{n} R(Y_1)$$

Tests were performed on actors across both a 5-year and 10-year interval. Additionally, a Spearman correlation will also be performed on the same data for the same actors and for the same intervals;

$$r_s = (r_s - E_0(r_s))/\{var_0(r_s)\}^{1/2}$$

For each hypothesis test the null hypothesis stated the absence of a relationship between the actor's cultural values and the actual observed behaviour. The following general hypotheses were examined.

> *H_{0s}: There is no statistical relationship between cyber source actors and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{1s}: There exists a statistical relationship between cyber source actors and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{0t}: There is no statistical relationship between cyber victims and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{1t}: There exists a statistical relationship between cyber victim actors and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{0s}: There is no statistical correlation between cyber source actors and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{1s}: There exists a statistical relationship between cyber source actors and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{0t}: There is no statistical relationship between cyber victims and cultural values when physical events have occurred. (Sample, 2015)*
>
> *H_{1t}: There exists a statistical correlation between cyber victim actors and cultural values when physical events have occurred. (Sample, 2015)*

4. Results figures and tables

Table 1: Mann-Whitney results for source actors

Dimension	p-value	Null Hypothesis	Alternative Hypothesis
PDI – 10 year	**0.0918**	**Consider both**	**Consider both**
PDI – 5 year	**0.0139**	**Reject**	**Accept**
IVC – 10 year	-0.3859	Accept	Reject
IVC – 5 year	-0.2148	Accept	Reject
M/F – 10 year	**-0.0107**	**Reject**	**Accept**
M/F – 5 year	**0.0262**	**Reject**	**Accept**
UAI – 10 year	0.2912	Accept	Reject
UAI – 5 year	0.3974	Accept	Reject
LvS – 10 year	-0.3520	Accept	Reject
LvS – 5 year	-0.1977	Accept	Reject
IVR – 10 year	**-0.0934**	**Consider**	**Consider**
IVR – 5 year	**-0.0188**	**Reject**	**Accept**

Table source (Sample, 2015)

Table 2: Mann-Whitney results for victim actors

Dimension	p-value	Null Hypothesis	Alternative Hypothesis
PDI – 10 year	0.1841	Accept	Reject
PDI – 5 year	0.1635	Accept	Reject
IVC – 10 year	0.2327	Accept	Reject
IVC – 5 year	0.1711	Accept	Reject
M/F – 10 year	**0.0143**	**Reject**	**Accept**
M/F – 5 year	**0.0233**	**Reject**	**Accept**
UAI – 10 year	-0.4052	Accept	Reject
UAI – 5 year	-0.2514	Accept	Reject
LvS – 10 year	-0.2514	Accept	Reject
LvS – 5 year	-0.0764	Accept	Reject
IVR – 10 year	-0.0778	Accept	Reject
IVR – 5 year	-0.1635	Accept	Reject

Table source (Sample, 2015)

Table 3: Spearman correlation results for source actors

Interval /Dimension	r	t	Correlation
PDI – 10 year	**0.6341**	**2.33**	**Strong +**
PDI – 5 year	**0.8**	**n/a**	**Strong +**
IVC – 10 year	-0.2188	-0.63	Weak -
IVC – 5 year	-0.7	n/a	Strong -
M/F – 10 year	0.1281	0.37	Weak +
M/F – 5 year	0.7	n/a	Strong +
UAI – 10 year	**-0.6**	**-2.12**	**Strong -**
UAI – 5 year	**-0.4**	**n/a**	**Moderate -**
LvS – 10 year	-0.2614	-0.77	Weak -
LvS – 5 year	-0.7	n/a	Strong -
IVR – 10 year	**-0.5801**	**-1.91**	**Strong -**
IVR – 5 year	**0.8721**	**n/a**	**Strong +**

Table source (Sample, 2015)

Table 4: Spearman correlation results for victim actors

Interval /Dimension	r	t	Correlation	Hypothesis
PDI – 10 year	0.2918	0.86	Weak +	Consider both
PDI – 5 year	**0.3**	**n/a**	**Moderate +**	**Consider both**
IVC – 10 year	-0.25	-0.73	Weak -	Accept null
IVC – 5 year	0.1	n/a	None	Accept null
M/F – 10 year	**-0.3643**	**-1.11**	**Moderate -**	**Consider both**
M/F – 5 year	-0.1	n/a	None	Consider both
UAI – 10 year	0.2683	0.79	Weak +	Accept null
UAI – 5 year	0.1	n/a	None	Accept null
LvS – 10 year	**-0.6322**	**-2.31**	**Strong -**	**Reject null**
LvS – 5 year	**-0.7**	**n/a**	**Strong -**	**Reject null**
IVR – 10 year	**-0.5495**	**-1.86**	**Strong -**	**Consider both**
IVR – 5 year	**-0.2**	**n/a**	**Weak -**	**Consider both**

Table source (Sample, 2015)

Source actors consistently showed results with high power distance, masculine and restrained traits. Source actors' uncertainty avoidance index also negatively correlated with the number of attacks. Restraint also appeared to negatively correlate with the number of attacks by source actors. The high power distance values positively correlated with source actor's number of attacks. Meanwhile, victims were masculine compared to the control group, while not being overtly masculine. Additionally, short-term orientation negatively correlated with the number of attacks.

5. Discussion

5.1 Attackers

An often-used method of defacement is via a Structured Query Language (SQL) injection. SQL-injection attacks results from an attacker entering requires the attacker to entering unexpected characters or inputs into a string that results in the program to make an unintended query to the database (Halfond & Orso, 2005; Su & Wasserman, 2006). The act of entering the data into the string is the injection component of the attack.

Successful SQL-injections allow attackers to gain administrative access into a website, or if the victims' username and password are uncovered first, then attackers can FTP inside, or enter by some other mean. No matter the method, defacement attackers must overtake a site, an action that is strongly aggressive and assertive, these traits that are commonly associated with high masculinity and high power distance. The firmness to burst in and gain control of victims' site is not only strongly masculine, but also bold, characterized by low Uncertainty Avoidance Index and high PDI values. Hofstede et al., (2010) observed that in a high PDI society might makes right (p.77) and in the low UAI countries that curiosity overrules fear (p.200).

Attackers who are intent on performing defacements utilize a strong sense of mission to deface the entire page, and many often leave tale tell signs, pseudonyms, codenames or insignia. Accomplishing the attack mission requires focus, care, and ability to deal with unforeseen complexities. In a way, attackers are showing victims how inferior or sloppy they are no matter how smart or vigilant they thought they were. This tendency to stay focused on task instead of give in to frivolous enjoyment is revealed by the restraint (as opposed to indulgence) pole of the IVR dimension. Another way to describe restrained countries is culturally tight, which can be defined as "strong norms and low tolerance of deviant behaviour" (Gelfand et al. 2011, p.1100). This echoes the observation by Hofstede (et. al., 2010) where the IVR dimension was linked to loose and tight societies (p. 281). The tight society will not only prohibit deviations from the societal norms but will also punish the perpetrator through the use of shame (Ibid). The act of publicly humiliating the masculine country for either perceived or real threats to the restrained culture's order may result in a need to shame the victim.

Finally, lack of ability to take on victims in the physical realm, where they often have unequally large measures of power and protection, expresses itself in clandestine take downs of victims' website covertly, confirming the power imbalance of the less-powerful individuals who have come to accept the hierarchical or uneven distribution of power. Real power may not available to attackers, but virtual cyber realms allow a perceived level playing field where the less powerful can bring the powerful to their knees by exerting their higher smarts and capabilities. Minkov's (2011, 2013) observations about the deeper understanding of math and sciences apply to the display of superior knowledge, suggesting an online version of the pen being mightier than the sword. The defacement activity itself prescribes the profile, qualities and requirements needed by attackers to accomplish such tasks successfully.

5.2 Victims

Entry into a website and defacement of the site is the attackers' way of humiliating the system administrators for failing to maintain server security. Many of the observed defacements displayed this, when the attacker left a message reminding the site owner that things could have been much worse and that they owner needs to learn security. Such an aggressive message denotes a high Masculinity dimension, but becomes all the more significant to attackers if the victims are highly masculine too, suggesting a need for the attacker to remind the victim who is "alpha" versus "beta". In effect, if the website belongs to individuals, organizations or nations that are iconic or highly masculine, then the act of defacement becomes even more effeminizing. Whether the defacement was meant to demonstrate the high skill level of the attackers, or sinister means of extracting valuable proprietary files, or uploading malware, the victims are temporarily brought down to their knees by the attackers.

Religious sites, government sites and corporations are often targeted by web defacers who are in effect either defacing the views or beliefs of others, the orderliness and governance of their society (or other societies), or the image and sense of reliability of the corporation under attack. Whether the defacements seek to convey a political or religious message or cause damage to the business of a corporation, the larger message the defacement conveys to victims is that they did not spend valuable time thinking ahead, taking important

precautions, buckling down to harden their web structures to withstand attackers' penetration schemes, to think through the weaknesses and vulnerabilities of their systems. These are all markers of the short-term orientation of the victims. Government owned website defacements communicate that the victim government is not doing a good job, is not reliable, is short-sighted, and/or incompetent.

The data on government owned websites confirmed that victims were masculine, yielding a more significant defacement success for the attackers, as well as short-term oriented, yielding the conclusion (to their public) that they were inattentive and sub-standard.

6. Conclusion

These findings support the findings by Henrich, Heine, and Norenzayan (2010) that universal behaviours cannot be claimed through sampling a single subpopulation. Considering the number of political disputes and the distribution of the actors in the kinetic realm showed no specific cultural markers, the findings on the patriotic, political defacements and cultural markers is significant. These findings apply to both victims and attackers. The recognition of the cyber domain as a conflict domain along with the knowledge that the cyber domain is unique, suggests that studies in this domain must fuse the human behavioural elements to a greater degree than what has been done in the past with the physical domains. This study along with other cultural studies by Sample (2013, 2015), Sample & Karamanian (2014, 2015) and Karamanian et al., (2016) bring the work of Hofstede et al., (2010), Nisbett (2010) and Henrich et al. (2010) along with others into the cyber domain.

The linkage between those who defaced government owned websites and culture was first established by Sample (2013). Subsequent studies by (Sample & Karamanian (2014), and Karamanian, Sample and Kolenko, 2016), added to the body of knowledge when they examined target (or victim) sites for cultural commonalities in vectors and platforms. Other commonalities were observed in messaging and naming, but were not studied due to scope and other limitations. The study of government website defacement data confirmed the predictions laid out by the Hofstede framework that defacers were more masculine, high in PDI, stronger in restraint (low IVR), and lower in UAI. The victims were masculine, and more short term oriented. The cultural markers for victim sites, while weaker than the attacking counterparts, were present.

References

Bargh, J.A., and Morsella, E., (2008) "The Unconscious Mind", *Perspectives on Psychological Science,* Volume 3, No. 1, pp.73-79.

Baumeister, R.F., and Masicampo, E.J., (2010) "Conscious Thought is for Facilitating Social and Cultural Interactions: How Mental Simulations Serve the Animal-Culture Interface", *Psychological Review,* Volume 117, No. 5, pp. 945-971.

Beidleman, S.W., 2009. *Defining and deterring cyber war*. ARMY WAR COLL CARLISLE BARRACKS PA.

Borkovich, D. J. (2012). When corporations collide: Information overload. *Issues in Information Systems*, *13*(2), 269-284.

Dijksterhuis, A. (2004). "Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making, *Journal of Personality and Social Psychology,* 2005, Volume 7, No.5, pp. 586-598. doi:10.1037/0022-3514.87.5.586.

Elmasry, M. H., Auter, P.J. & Peuchaud, S. R. (2014). Facebook across cultures: A cross-cultural content analysis of Egyptian, Qatari and American student Facebook pages. *Journal of Middle East Media*, vol.10, fall 2014.

Gelfand, M.J., Raver, J.L., Nishii, L., Leslie, L.M., Lun, J., Lim, B.C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J. and Aycan, Z., 2011. Differences between tight and loose cultures: A 33-nation study. *science*, *332*(6033), pp.1100-1104.

Gonzalez, C. and Dutt, V., 2011. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological review*, *118*(4), p.523.

González-Vallejo, C., Lassiter, G.D., Bellezza, F.S. and Lindberg, M.J., 2008. " Save angels perhaps": A critical examination of unconscious thought theory and the deliberation-without-attention effect. *Review of General Psychology*, *12*(3), p.282.

Halfond, W. G., & Orso, A. (2005, November). AMNESIA: analysis and monitoring for neutralizing SQL-injection attacks. In *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering* (pp. 174-183). ACM.

Hayden, Michael V. "The future of things CYBER." *Conflict and Cooperation in Cyberspace: The Challenge to National Security* (2013): 1.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33, 61-83.

Hofstede, G. (2011). "Dimensionalizing Cultures: the Hofstede Model in Context", *Online Readings in Psychology and Culture,* Volume *2,* No. 1, p. 8*.*

Hofstede, G., Hofstede, G.J., and Minkov, M. (2010). *Cultures and Organizations,* McGraw-Hill Publishing: New York, NY.

Internet World Stats website (2013) www.internetworldstats.com.

Karamanian, A., Sample, C., and Kolenko, M. (2016). "Hofstede's cultural markers in successful victim cyber exploitations", *Proceedings of The 11th International Conference on Cyber Warfare and Security ICCWS 2016",* Boston University, Boston, MA, March 17-18 2016, 205 -213.

Kaur, A. & Noman, M. (2015). Exploring classroom practices in collectivist cultures though the lens of Hofstede's model. *The Qualitative Report*, 20(11), 1794-1811. Retrieved from: http://nsuworks.nova.edu/tqr/vol20/iss11/7

Khanaki, H., & Hassanzadeh, N. (2010). Conflict management styles: The Iranian general preference compared to the Swedish. *International Journal of Innovation, Management and Technology*, *1*(4), 419.

Liao, Q.Y., Pan, Y.X. & Zhou, M. (2009). Chinese online communities: Balancing management control and individual autonomy. *IBM Research Report*. Retrieved online: http://domino.research.ibm.com/library/cyberdig.nsf/papers/332978E338D8446D8525768F005F235A/$File/rc24919.pdf

Minkov, M. (2011). *Cultural Differences in a Globalizing World.* WA, UK: Emerald Group Publishing Limited.

Minkov, M. (2013). *Cross-Cultural Analysis.* Thousand Oaks, CA: Sage Publications.

Morris, M.W. and Peng, K., 1994. Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social psychology*, *67*(6), p.949.

Nisbett, R., 2010. *The Geography of Thought: How Asians and Westerners Think Differently... and Why*. Simon and Schuster.

Sample, C., (2013). "Applicability of Cultural Markers in Computer Network Attack Attribution", *Proceedings of the 12th European Conference on Information Warfare and Security,* University of Jyvaskyla, Finland, July 11-2, 2013, pp. 361-369.

Sample, C. (2015). "Cyber + Culture Early Warning Study" CMU/SEI-2015–SR-025, (2015) retrieved from: http://resources.sei.cmu.edu/asset_files/SpecialReport/2015_003_001_449739.pdf

Sample, C. and Ara Karamanian (2014). "Hofstede's Cultural Markers in Computer Network Attack Behaviours", *Proceedings of the 9th International Conference on Cyber Warfare and Security, ICCWS 2014,* Purdue University, West Lafayette, Indiana USA, March 24-25, 2014, pp. 191 – 200.

Sample, C., and Karamanian, A., (2015, July). Culture and Cyber Behaviours: DNS Defending. In *Proceedings of the 14th European Conference on Cyber Warfare and Security 2015: ECCWS 2015* (p. 233). Academic Conferences Limited.

Shimizu, M., Park, H & Greenfield, P. M. (2014). Infant sleeping arrangement and cultural values among contemporary Japanese. Frontiers in Psychology, vol.6, p. 264, August 2014; doi: 10.3389/fpsyg.2014.00718. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137277/

Su, Z., and Wassermann, G. (2006, January). The essence of command injection attacks in web applications. In *ACM SIGPLAN Notices* (Vol. 41, No. 1, pp. 372-382). ACM.

Wei, W. U. (2001). *Individualism-collectivism and Conflict Resolution Styles: A cross-cultural study of managers in Singapore* (Doctoral dissertation, Department of English, City University of Hong Kong).

Zone-h website (2012). http://www.zone-h.org.

Utilising Journey Mapping and Crime Scripting to Combat Cyber Crime

Tiia Somer[1], Bil Hallaq[2] and Tim Watson[2]
[1]Tallinn University of Technology, Estonia
[2]University of Warwick, UK
tiia.somer@ttu.ee
bh@warwick.ac.uk
tw@warwick.ac.uk

Abstract: Modern society is now reliant on digital communication and networks for conducting a wide array of tasks, ranging from simple acts such as browsing the web through to mission critical tasks such as the management of critical infrastructure and industrial controls. This reliance shows a growing emphasis on strategic importance of cyberspace (Sharma, 2010). While organisations and individuals are keenly exploiting the benefits of cyberspace, these same platforms have also opened new avenues for nefarious actors in the pursuit of their criminal activities to attack, disrupt, or steal from organisations and individuals. Criminal organisations and lone criminals worldwide have access to powerful, evolving capabilities which they use to identify and target their victims allowing for the perpetration of a wide variety of cyber crimes. This paper discusses ways in which utilising methods from typically non-cyber disciplines – business and criminology – can successfully be applied to the cyber domain in order to help in the fight against and prevention of cyber crime. Through the provision of a visual representation, this paper clarifies how journey mapping and crime scripting can help in building an understanding of the steps criminals undertake during execution of a cyber crime. In essence, within our work we have deconstructed the lifecycle of a crime events and translated these into a visualisation map to show the full event process, highlighting key steps as well as positive and negative events. Such work is useful to several roles and organisation types as it can aid in their decision processes when undertaking steps in pursuit, prevention, preparation and protection.

Keywords: cyber crime, criminal journey mapping, cyber crime scripting, cyber crime pathways, E-CRIME project

1. Introduction

It is an established fact that the internet has greatly affected the way societies and people operate. Worldwide internet usage has increased to more than 3.5 billion users at the beginning of 2014 (Internet Usage and World Population Statistics, 2014). In addition to people, internet connectivity today extends to digital devices, with more things connected to the Internet than people. Gartner predicts that the number of internet connected devices will reach 25 billion for 2020 (Gartner, 2014).

While organisations and individuals are quick to exploit the business and personal benefits of internet, they often give less consideration that cyberspace offers a plethora of benefits to those who wish to attack them. Hacker groups, criminal organisations and espionage units worldwide have access to powerful, evolving capabilities, which they use to identify and target their victims and commit cyber crimes.

This research was conducted as part of the Economic Impacts of Cybercrime (E-CRIME) project of the Seventh Framework Programme, funded by the European Union. The majority of this work has been conducted with the help of desktop research and insights from a group of experts; the conclusions drawn and statements made rely on the Deliverable 2.3. “Detailed appendixes on cyber crime inventory and networks in non-ICT sectors” of the E-CRIME project. The work was conducted by means of a review of the existing literature and an evaluation of the published approaches, as well as by conducting expert interviews. Sources of information included journals and conference proceedings in the fields of law, criminology and information systems, reports published by think-tanks and law enforcement agencies as well as scholarly textbooks.

Interviews of experts were also undertaken as a further means of data collection with the main consideration being: even though cyber crime has been researched extensively, the specific criminal “journeys” and stepping stones the cyber criminals take within crime cycles have not been subject to such research methods previously to the best of the authors knowledge based on publically available information. An interview guide was prepared, which provided an informal grouping of topics to be covered during the interview. Once completed the results of the interviews provided extra data and some interesting nuances. The authors prioritized the interview results, since the main focus was the provision and mapping of criminal journeys.

The expert groups of interviews for this paper consisted of law enforcement operating at regional, national and international levels, industry based cyber security experts as well as experts from academia. The aim was to reach a common conclusion, and not to research single activities at the micro levels. Different focus groups each had specific expertise and points of view to the topic of the research – cyber crime – which with the method chosen allowed for analysis of the experiences and requirements of a wider audience.

2. Cyber crime

Cyber crime is increasing in both complexity and intensity, reflecting an increased level of sophistication. For the purposes of this work our focus on cyber crime includes different aspects and extensions of modern crime: from development and sale of attack tools, services to plan and execute attacks and culmination in the laundering of stolen or illegally obtained assets. Cyber criminals increasingly operate in the same manner as legitimate business networks with clearly established business objectives and trusted supply chains for services or products that require outsourcing or development. The cyber criminals know what they are looking for, what goals they want to achieve and how to achieve these goals – and they are willing to spend time to research and plan their actions (CISCO 2014).

Given the complex nature of cyber crime and in order to understand and take efficient measures against it, it is imperative to gain deep understanding of the mechanics of cyber crime, from preparation, or pre-crime stages, to exit strategies and monetization, including everything in between the two including the committing of the actual crime. For the purposes of this paper we have performed several journey mapping exercises to describe the events and experiences that cyber crime perpetrators go through during a crime, using crime scripting techniques as found in traditional "offline" criminology. The research focus of this work has been on the crime itself, not the underlying causes of crime or the law enforcement actions following the crime. This mapping will help facilitate identification and testing of effective countermeasures, as well as facilitate further work in identification of possibilities to deter criminals and manage risks deriving from the perpetration of cyber criminal activities.

Three phases are critical to the development of our journeys from the perspective of the criminal:

- 1. Preparation phase
- *Decision to engage in criminal activity*
- *Choosing a victim*
- *Choosing a method*
- 2. Execution phase
- *Conducting the crime*
- 3. Monetization/Reward phase
- *Exit*

Cyber crime can be seen as a process where resources are required and decisions are taken at different stages in the process. The preparation phase includes pre-attack actions including committing to the initial decision to undertake a crime, deciding on the worthiness of an attack, identifying potential victims, and conducting targeted reconnaissance, but also a choice of an attack method including use of own means and abilities, or taking the decision to outsource respective capabilities. The execution phase includes drawing an attack plan and executing the attack itself, including entering the target system and conducting criminal activities within such systems. It also includes lateral movement and finding additional opportunities for criminal action. The monetization phase includes direct or indirect monetary gain for the cyber criminals(s) and exit strategy. It is important to note that throughout any one criminal journey, the perpetrator can loop back to an earlier step (if a chosen attack method fails, they need to find a new one, or they may 'accidentally' find unforeseen vulnerabilities to take advantage of), or they can repeat steps for example, defacing the same website multiple times, or they may just quit once they realise the efforts are not worth the results.

Various sources show the developments of global cybercrime and related threat landscape. The United Nations Comprehensive Study on Cybercrime of 2013 states that cybercrime globally shows a broad distribution across financially driven acts, computer-content related acts, but also attacks against the confidentiality, integrity and availability of data and computer systems (United Nations 2013) which is key to take into consideration in

understanding cyber criminal journeys. The 2015 RSA outlook on the changing threat landscape of cybercrime states that the most important trend developing within the past few years, is the rapid advancement of cybercrime-as-a-service model. What this development means, is that more criminals can participate in the chain and that these criminal do not need to understand the complete chain of the crime nor how to conduct any specific part of it, for example spam, DDoS or phishing. Nor do they need to have the technical requirements in house to the conduct of the crime itself (RSA 2015). The ENISA Threat Landscape 2015 states that from cybercrime-as-a-service model, the most mature are botnet-related service models (ENISA 2015). ENISA also states that the most rapidly growing service is provision of ransomware-related services. These points clarify the importance of journey mapping and crime scripting in order to provide those combatting cyber crime with a clear understanding of the complete crime cycle, including the various aspects and actors which may take part at different phases of a cyber crime.

3. Journey mapping

Journey mapping is a methodological tool that has been traditionally used in business to map customer experience, as well as in criminology generally under the name of crime scripts. Journey mapping is also often used by strategy consultancies and public organisations to shape customer strategies and public service transformational programmes. In criminology, crime scripts have been used to deconstruct complex crimes into component parts even from a relatively small data set. Within this work we have used such methods from these typically non-cyber disciplines and shown that they can be successfully applied to the cyber domain.

4. Crime scripting

The 'map'-style of output has been adopted and applied within a number of different disciplines where it is often referred to as a *script*. A script is a predetermined set of actions that define a well-known situation in a particular context (Borrion, 2013), or more specifically "[a] script is simply a sequence of actions which make up an event" (Brayley, 2011). Scripts are related to the concept of schema, i.e. "abstract cognitive representations of organised prior knowledge, extracted from experiences with specific instances". When the sequence of events being scripted encapsulates the conduct of a criminal activity (as in the case of cyber crime), the output is commonly referred to as a crime script (Borrion, 2013). Initially developed in psychology, scripts are now used in different fields from artificial intelligence to consulting.

Scripts can be used to present different crimes, but are believed to be of particular use for new or complex crimes (Brayley, 2011). It has also been suggested that crime scripts can be used as an innovative way to gain a more detailed understanding of complex forms of crime in a review of organized crime-reduction strategies (Levi, 2004). As previously stated in this paper, cyber crime is a rapidly developing field with an evolving trend of the cybercrime-as-a-service model. This will bring more participants into the cyber crime journey or cycle, making it more complex to understand for those dedicated to prevention and fight against cyber crime.

5. Why can criminal journey maps be useful?

By schematically representing an anticipated sequence of actions, scripts are able to provide us with a cognitive representation of how we believe a sequence of events has occurred and will occur (Borrion, 2013), including for our purposes, the steps a criminal takes to commit a cyber crime. In this situation, the value of crime scripting as a crime analysis mechanism is believed to be in its potential to assist in the fight against such crime (Borrion, 2013) through the identification of *pinch points*. For example, by graphically presenting the typical sequence of events for a crime that has been derived from many examples of that type of crime, analysts are able to identify specific metaphorical gates the criminal must pass through if their crimes are to succeed. Once these points are identified, the logic is that those seeking to prevent such crimes will now know where best to focus their energies, whether this be through legislative or regulatory changes, the development of new technological countermeasures, development of general awareness campaigns, the behaviour change of potential victims, or increased monitoring by police forces so as to capture or deter the cyber criminals. As stated in many cybercrime related sources, cyber crimes are becoming more complex, involving more parties each conducting independent steps within various phases of any one crime (RSA 2015, ENISA 2015). The understanding of each step, however minor within this crime cycle, will become more vital. The journey maps developed provide a cognitive representation of how we believe a cyber crime takes place from preparation to monetization and exit.

Some crime scripts list a sequence of actions and don't draw a diagram, others draw a graphical representation showing a series of actions and decision points. In graphical presentations, scripts are usually drawn as series of

boxes, linked by arrows indicating direction of flow (where boxes indicate actions or decisions). As the same crime can be committed in different ways, so can different routes/tracks co-exist on one script.

There are various levels of scripts and selection depends on the script's intended application (Brayley, 2011). For the purposes of the current work, we developed a high-level journey map detailing a general cyber crime cycle (Figure 1). This is a general depiction of a single cyber crime act, from which more detailed maps in different categories can been drawn. In order to be of practical use in understanding cyber crime, more detailed journey maps for different criminal journeys are needed, providing crime sequences from preparation to exit for these specific journeys.

Since there are no standard journey mapping rules or specific software for crime scripting (Brayley, 2011), we have used our own symbols and drawings. We grouped similar actions under broad terms: preparation, execution, and monetization. The journey maps developed provide a step-by-step high-level account of actions taken by the criminals throughout the crime. Crimes are a process which involves several steps leading to reaching an end-goal as identified by respective criminals. For example, the preparation phase includes various pre-attack actions, i.e. initial decision, deciding the worthiness of an attack, identifying victims, and conducting targeted reconnaissance. The preparation phase also includes the choice of an attack method, including the cyber criminal(s) undertaking an analysis of their own means and abilities and making the decision of outsourcing or buying solutions from external sources in case there is a resource or skills gap. The execution phase includes creating an attack plan and executing the attack, which comprises of entering or interfacing with target system and the actual criminal activities (i.e. distributed denial of service (DDoS), extortion, espionage, etc.) themselves. However, it is important to note that the tactics used by criminals do not always follow the above formalised decision points, meaning that in some instances decisions are made very quickly without conducting a full-scale analysis or creating a set of actual attack plans. A further important point to note, is that the criminal can loop back to any earlier phase as required by circumstances and in some instances they may choose to abort the undertaking for example in cases where the criminals might determine it is no longer cost-effective or the potential risk of getting caught is not worth the reward. The monetization phase includes a tangible payment in some form with laundering and/or mules often being utilised, although in some instances the criminals will not have a monetary objective which is discussed in the next section. The final result culminates in a personal gain or fulfilment of end-results as set out in the initial stages for the criminal(s).

6. Mapping and scripting a general cyber crime journey

Figure 1 represents a high-level journey map detailing a general crime cycle from the criminal's perspective. This general cyber criminal journey and journeys for any follow-on specific crimes have been developed with the help of desktop research and insights from experts as part of the E-Crime 7th framework project as already mentioned in this paper. The benefits for investigators of producing this visual representation of the general cycle are that;

- By identifying the commonalities in the conduct of what may seem very different cyber crimes, we can expose the sequence of events that underpin the majority of these.
- By comparing detailed maps of multiple different cyber crimes against this general crime cycle, those tasked with preventing and/ or defending against such crimes can see best where to focus their resources for maximum effect.
- Experiment via virtualised or desktop exercises the application of countermeasures at various points along the pathway with the goal of taking forward the most effective ones for application in real word scenarios.

The preparation phase has two main components. Firstly the criminals need to decide whether or not to conduct the crime in the first place. This may be simply an opportunistic decision or it may include "market research" of some kind, in the sense of determining and weighing the costs and benefits of their options. The second component requires the identification of potential victims and attack methods, the conducting of targeted reconnaissance and finally deciding to execute the criminal act. The attack itself can then be executed in three ways, either; (1) by using their own existing means and abilities. As an example in case they have access to their own botnet, malware or exploit already, they will use these and will not go through all the steps in the process but move to the execution phase directly. (2) By buying the respective means and/or capabilities from other criminals – this brings other sectors into the process (special markets, forums, stores, sellers, brokers, developers, etc.), or (3) Outsourcing the required criminal activities by paying another criminal to conduct it as a service (crime-as-a-service).

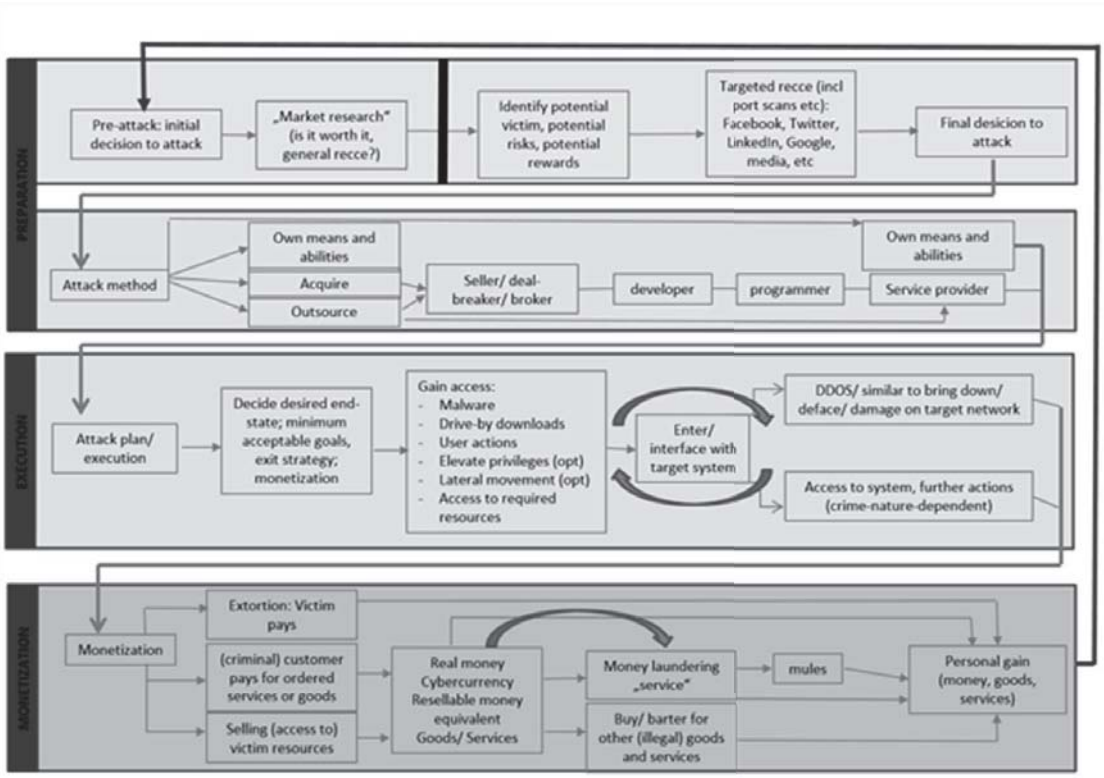


Figure 1: General cyber crime journey map

The execution phase starts with an attack plan. In the plan the criminal decides upon a desired end-state, their minimum acceptable goals, and monetisation and exit strategies. During the attack, the criminal gains access to victim's resources through any number of means, including malware, drive-by downloads, user actions via phishing techniques or other illicit activities. Once the criminal gains access to the victim's system, they will map the compromised network, often looking for further opportunities to exploit. Thereafter the criminal enters or interfaces with the target system and based on their desired and decided goals and end-states they take the commensurate actions. Within this phase of mapping the compromised network, the criminal may notice other vulnerabilities that may become useful in their reaching of stated end-results and will take advantage of these, i.e. committing different crimes which were not originally part of their attack plan.

The monetization phase involves obtaining tangible benefits. These benefits include direct monetary gain, for example where the victim's monetary assets are stolen, or the victim pays the criminal directly in cases of extortion, such as ransomware or DDoS extortion schemes. Or indirect monetary gain whereby the victim's resources can be turned to tangible assets which are traded or sold, for example selling access to the victim's machine to others. The payment can be conducted in real currency, crypto-currency, resalable money equivalents (such as gaming assets), or in goods and services (real or virtual, legal or illegal). In some cases money laundering services are used, in other cases other means such as setting up mules to withdraw cash from banks might be used. Clearly though in some cases the monetization phase is excluded, examples of such cases include Hacktivists or those with ideological or other motivations.

In any case the crime ends with an exit strategy as set out by the criminal culminating in some type of personal gratification be it monetary or otherwise.

7. Conclusion

Within this report, the authors have shown how traditional crime scripting can provide useful insights into understanding the lifecycle of a general cyber-criminal journey. It also shows how methods and techniques from typically non-cyber disciplines can be successfully applied to the cyber domain. Such mapping and scripting can be modified and further detailed for specific crime scenarios and graphically represented. By graphically presenting the sequence of events constituting a cyber crime, risk management teams, forensic analysts, incident response teams and law enforcement agencies will be able to identify the specific stepping stones and pinch points that cyber criminals pass through in committing their crimes. Such work can help to facilitate the

identification and testing of effective countermeasures including mitigation at scale, early prevention and the development of proportional disruption techniques.

References

Brayley, H., Cockbain, E., Laycock, G., 2011. The value of crime scripting: Deconstructing Internal Child Sex Trafficking, Policing, Volume 5, Number 2, pp. 132–143

Borrion, H., 2013. Quality assurance in crime scripting, Crime Science 2013, 2:6. Available online at: http://www.crimesciencejournal.com/content/2/1/6

CISCO, 2014. Annual Security Report. Available online at: http://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2014_ASR.pdf

The Economic Impacts of Cyber Crime, FP7-SEC-2013.2.5-2. D2.3 Detailed appendixes on cyber crime inventory and networks in non-ICT sectors. T.Sõmer, R.Ottis, T.Lepik, M.Lagazio, B.Hallaq, D.Simms, T.Mitchener-Nissen. March 2015

ENISA Threat Landscape 2015. ENISA 2015. Available for download at: https://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-threat-landscape/etl2015

Gartner, 2014. http://www.gartner.com/newsroom/id/2905717

Internet Usage and World Population Statistics, 2015. http://www.internetworldstats.com/stats.htm

RSA 2015. CYBERCRIME 2015: An Inside Look at the Changing Threat Landscape. Available online at: https://www.emc.com/collateral/white-paper/rsa-white-paper-cybercrime-trends-2015.pdf

Sharma, Amit, "Cyber Wars: A Paradigm Shift from Means to Ends", Strategic Analysis, Vol. 34, No. 1, 2010, pp. 62-73. http://www.tandfonline.com/toc/rsan20/34/1

United Nations 2013. United Nations Office on Drugs and Crime, Comprehensive Study on Cybercrime, February 2013. Available online at: <http://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf

Extracting Intelligence From Digital Forensic Artefacts

Stilianos Vidalis[1], Olga Angelopoulou[1] and Andy Jones[2]
[1]Cyber Security Centre, School of Computer Science, University of Hertfordshire, UK
[2]Security Research Institute, Edith Cowan University, Australia
s.vidalis@herts.ac.uk
o.angelopoulou@herts.ac.uk
a.jones26@herts.ac.uk

Abstract: Forensic science and in particular digital forensics as a business process has predominantly been focusing on generating evidence for court proceedings. It is argued that in today's socially-driven, knowledge-centric, virtual-computing era, this is not resource effective. In past cases it has been discovered retrospectively that the necessary information for a successful identification and extraction of evidence was previously available in a database or within previously analysed files. Such evidence could have been proactively used in order to solve a particular case, a number of linked cases or to better understand the criminal activity as a whole. This paper will present a conceptual architecture for a distributed system that will allow forensic analysts to forensically fuse and semantically analyse digital evidence for the extraction of intelligence that could lead to the accumulation of knowledge necessary for a successful prosecution.

Keywords: intelligence-led policing, evidence fusion and dissemination, forensic intelligence, ID theft

1. Setting the scene

A few years ago, a case was brought to a successful conclusion. The suspect was convicted for a number of computer-related crimes based on digital evidence that were extracted from computing devices found in his possession, following the standard and very well published dead-box digital forensics analytical procedure. After the end of the proceedings, one of the authors was given access to the evidence for further analysis. Such analysis was previously considered outside the scope of the investigation. The author was then able to extract actionable intelligence, linking the convict to a more serious crime and a number of other criminal activities on a different continent, committed in collaboration with a number of foreign nationals.

In the past, computer-related crimes were defined as those activities where computers were used for the commission of crime, where computers contained evidence of crime and/or where computers were the targets of crime (Hale 2002). Given today's (socially-driven knowledge-centric virtual-computing era) specific parameters, there is a need for a slightly different and more inclusive definition, addressing the different types of computing devices, e.g. mobile phones, smart embedded devices, game consoles, laptops, computers, etc. and a domain that goes beyond the concept of the term "cyber-domain". For the purposes of our research we will use the term Information Environment (IE). The U.S. Department of Defence (DOD) has defined the Information Environment (IE) in its Joint Publication 3-13 for Information Operations (US DoD 2012), stating that *"... the information environment is the aggregate of individuals, organizations and systems (resources) that collect, process, disseminate, or act on information."* Hence, computer-related crimes can be defined as:

> *Activities where physical and logical computing devices, attached to an Information Environment, are used for the commission of crime, contain evidence of crime, and/or are the targets of crime.*

Continuing our reasoning, and in agreement with the Association of Chief Police Officers (ACPO) guidelines, digital evidence, or computer-based electronic evidence is information and data of investigative value that is stored on or transmitted by a computer. Menou (1995, as cited in Chowdhury and Vidalis, 2013) described information as "*a product, which encompasses information as thing, as object, as resource, as commodity, what is carried in a channel (including the channel itself), the contents*."

Combining all of the above definitions, and accepting that information is paramount to the resolution of any crime, we can also accept that only having forensic evidence is no longer adequate for resolving computer-related criminal activities. Today, investigators need to have forensic intelligence (Ribaux et.al. 2003), (Legrand and Vogel 2012), even for the simplest and most trivial computer-related crime, that can lead to forensic evidence which, when combined, can lead to a strong supporting case for a prosecution. Such intelligence can be used either in a pro-active or in a re-active manner. As a concept, this is not new. It was first introduced and discussed a number of decades ago (Birkett 1989), (Ribaux and Margot 1999). For example, in the UK, ENDORSE (National Crime Agency 2015) is a nation-wide forensic and law enforcement initiative to collect and analyse

information from drug seizures made in the UK. Apropos, the use case for ENDORSE is limited to a specific problem and a specific crime type within one national jurisdiction. Furthermore, computer-related criminal activities can be seen as a very complex problem, combining different types of traditional criminal activities with different and innovative technologies for transcending jurisdictional boundaries.

Intelligence is the timely, accurate and usable product extracted from logically processed information. For this extraction to be successful and accurate, one must apply specialist knowledge. In the case of forensic intelligence, the concept of knowledge is twofold:

- procedural knowledge on the identification, individualisation, association and reconstruction of forensic evidence, and
- crime-specific analytical knowledge for translating leads to actionable forensic intelligence.

Before discussing the requirements for a system able to handle forensic intelligence, it is considered beneficial to identify problems with the current practice of resolving computer-related criminal activities.

2. Operational level issues

It is indicated (Statistic Brain 2015), (Mkomo 2015) that the cost of storage per GB of data is currently $0.03. Cloud storage is even cheaper according to The Register (2014). The cost of undertaking criminal activities is coming down. Even worse, because of the latest computing innovations such as virtualisation, cloud applications, communication applications and connectivity and interconnectivity opportunities, the physical crime scene is often different from the logical cybercrime scene.

According to Statista.com, 364.59 million Hard Disk Drives (HDDs) shipped globally in the first three quarters of 2015, and a figure of 416.7 million HDDs and 153.8 million Solid State Drives (SSDs) was projected for the whole of 2015. According to Digitaltends.com, the average size of the Seagate HDDs is now over one terabyte. Based on these statistics, we can assume that a typical case would require Law Enforcement Agency (LEA) Officers to collect, on average, more than 1TB of data (including CDs, DVDs, internal and external HDDs/SSDs). The automated procedures that can be used to assist in the processing of this data, such as file signature analysis and hash analysis, are employed. Apropos, a large amount of data has to be manually analysed. Even before the analysis stage, there is a lot of work to be undertaken. As part of a testing activity in a digital forensics laboratory, the authors had to clean a hard disk. Forensically wiping one Samsung HD105SI 1TB drive, using a tableau TD2u, was achieving an average of a 6.6GB/min transfer rate and a projected turnaround time of 2h 30 minutes. Furthermore, in a recent disk study the authors performed, a large number of hard disks were acquired and forensically analysed. The average acquisition transfer rate that was achieved was 2.76GB/min. This translates on an average time investigators would need to spend in the acquisition phase of at least 6 hours per disk.

After the acquisition of the devices, a forensic analyst will get to the analysis phase, where, depending on the case, they will perform any/all of the following activities:

- Disk geometry analysis (number, size and type of partitions (deleted or not))
- Time-zone analysis
- Operating System analysis
- Hash analysis
- File signature analysis
- Registry analysis
- Compound file analysis
- Log file analysis
- Internet artefacts analysis
- Email analysis

Following the above, more specific analytical steps will have to be performed (the list is not meant to be comprehensive):

- Deleted files recovery

- Identification of USB devices that were ever connected and when they were connected
- Identification of files and folders that have been exfiltrated
- CD/DVDS that may have been burned
- Websites visited and by which user account
- Lists of recently used programs, the files they have accessed, and when they have done so
- Programs that have been installed and uninstalled
- Attempts at data destruction/hiding
- Program settings that can deduce knowledge of an act or technology
- What programs start when the computer starts and any related DLLs, cross-examining findings for the identification of malware footprints
- How many times a program has ever been run and by which user account
- Wi-Fi connection points that have been accessed and when
- Hidden email and other internet accounts
- Identify and analyse photos and deduce the geographic location of where photos were taken
- What particular user performed a task (related to the above activities or to case specific activities)

Nowadays, most of the above analytical tasks have been automated. Still, depending on the datasets used, the analysis phase will take on average at least two days per disk to complete. This translates in two days per disk before the forensic analyst will be able to start the manual analytical activities, the file indexing and any case-specific raw searches. It also translates in two days that physical computing resources will have to be locked down and assigned to the execution of the aforementioned tasks.

Operationally, all of the above issues have a significant impact and create backlogs that, over time, can become unmanageable. An Open Source Intelligence (OSINT) search conducted in February 2015 by the authors indicated that police forces in the UK have backlogs that range from anywhere between 12 to 24 months. To overcome this issue, non-specialist staff are deployed, using of-the-self triaging products to minimise the collected artefacts, and even then, often the evidence is split between different analysts, with different levels of experience, in different teams, following slightly different analytical SOPs. Even worse, often LEA analysts stop their analysis once they have identified/extracted enough evidence to support an argument for a successful prosecution.

Our hypothesis is that analysts can make use of forensic intelligence, which once combined with forensic evidence can streamline and optimise the analytical process in a cost effective and appropriate (from a penology perspective) manner.

3. Requirements for fusing and semantically analysing evidence

In order to support our hypothesis, we will use identity theft as an example cyber-crime type. Identity theft is one of the major concerns today (CIFAS 2015), (ITRC 2015), (Harell 2015). It has a significant human component and is being strongly influenced by the way people treat personal information (defined and discussed in the doctoral thesis of Dr. Angelopoulou). BBC, amongst other reputable online sources, has published that in the first quarter of 2015 the number of ID-theft victims rose by 31% (BBC 2015). These statistics suggest that it is extremely difficult to eliminate identity theft by employing stronger computer security techniques since the perpetrators constantly find ways around them. Furthermore, based a report from Experian (2015) individuals often do not adopt any measures to protect themselves online. The average person easily provides and shares personal identity information relating to one or more of their online aliases.

> *"Regard your good name as the richest jewel you can possibly be possessed of - for credit is like fire; when once you have kindled it you may easily preserve it, but if you once extinguish it, you will find it an arduous task to rekindle it again. The way to gain a good reputation is to endeavour to be what you desire to appear". Socrates (469BC-399BC)*

Personal identity information (PII) is increasingly being stored and used in a range of digital forms. If this information is not adequately protected, this can leave individuals exposed to a range of possible threats.

Identity Theft (ID theft) is defined as someone's action of using any sort of distinct personal private information with fraudulent intention; mainly for financial gain (Angelopoulou et.al 2007). Technology related examples include; identity theft malware and key loggers, phishing, web-spoofing, online social engineering and database data retrieval. The complexity of retrieving substantial evidential information of an ID theft crime perpetrated over the Internet demands a specific methodological instrument that will be able to identify and extract evidential components (forensic intelligence) from different cases (in a big data analysis context). When such incidents are examined in detail, then the nature of the problem can be more clearly understood.

For our example we are utilising a platform that has the ability to extract, fuse and share ID-theft related forensic intelligence. The Standardised Forensic Intelligence Platform (SFIP), which is described in the following sections, runs the Identity Theft Investigations (ITI) module to extract, fuse and share ID-theft related forensic intelligence. The intelligence data will be hosted on a unified database from which LEA Officers will be able to collect case related knowledge relating to the suspect, the victim and the modus operandi MO, while waiting for the digital forensic analytical activities to conclude. The use case for our proposed system is presented in figure 1.

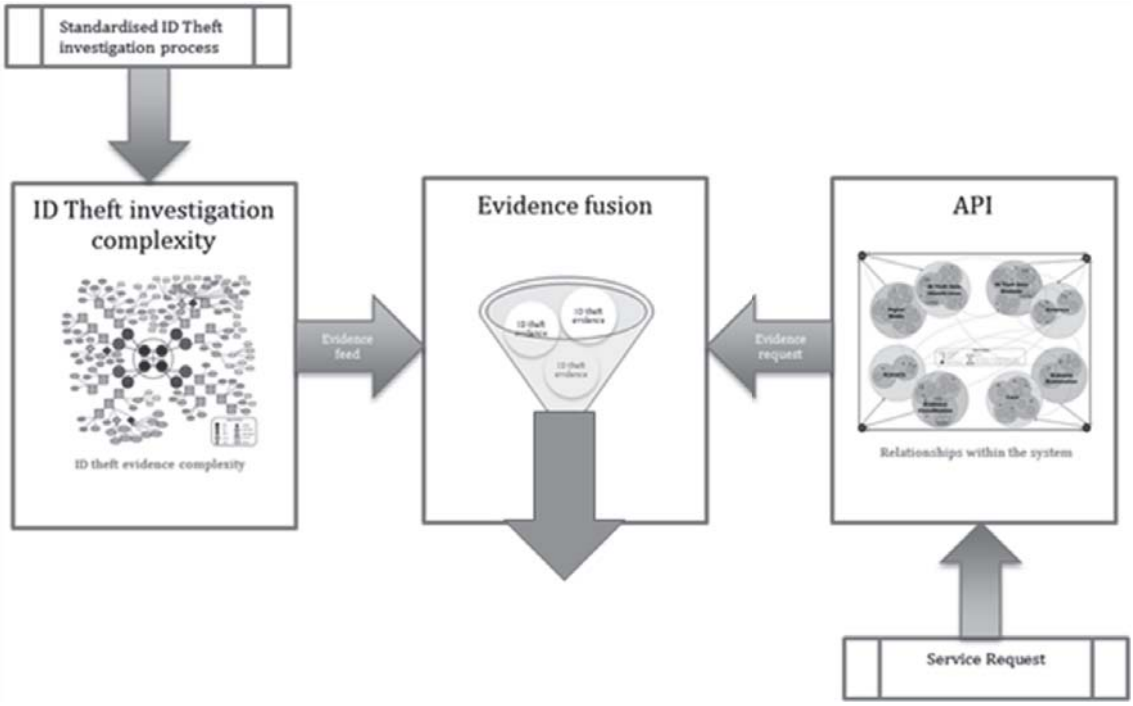


Figure 1: SFIP use case

The intelligence, knowledge and expertise will result from classified evidence that will be produced after a certain process and analysis has been followed. The fusion and semantic analysis of the evidence should provide a systematic application to a digital investigation of a cybercrime. We have summarised the requirements such a system should adhere to in the following list:

- Locally deployed modules must be connected to the local analytical workstations running different toolkits from different vendors. There is a requirement for a plug-in mechanism to allow for the development of new application program interfaces (APIs) as new forensic toolkits come to the market.
- The local database holding the intelligence data must replicate the manner in which data are being recorded by the Forces and obviously comply with ISO2700.
- Nodes must communicate in a secure manner. Encryption of all communication channels and of all messages (multi-layered encryption).
- Authentication and validation of all nodes and users.
- Signatures for providing non-repudiation and message integrity for all communications between system stakeholders/users/actors
- Logging of user and system requests and activities to ensure and assure chain of custody

- A metadata distribution system to ensure and assure evidential integrity and validity/authenticity of assets/artefacts.

4. Conceptual architecture

The proposed system will guide the forensic analyst through the intelligence collection process as the standard operating procedure (SOP) steps/activities will be built into the system. In our example as illustrated in figures 1 and 2, we are using the ITI module and an id-theft crime specific SOP. Semantic analysis using crime specific ontologies will be employed in order to extract crime-specific intelligence. Semantic techniques will enable contextual and relevant data to be identified for a particular entity. The use of ontologies will create a bridging mechanism, whereby semantic metadata could be referenced and validated to ensure that relevant and useful information is collected. This also ensures that trust and logic can be attained in the service request functionality. In the ITI module for example, the extracted knowledge (from the intelligence data) will link devices and content from different cases together, providing investigators with leads and forensic evidence that can be used for the profiling and prosecution of perpetrators.

SFIP will effectively present next to real-time knowledgeable answers to runtime user generated queries. It will collect information from disparate sources and use semantics to safeguard the future of knowledge discovery and reuse. The stakeholders will be able to request access to SFIP through an application program interface (API) and query specific relationships within the system.

The conceptual architecture of the system is illustrated in figure 2.

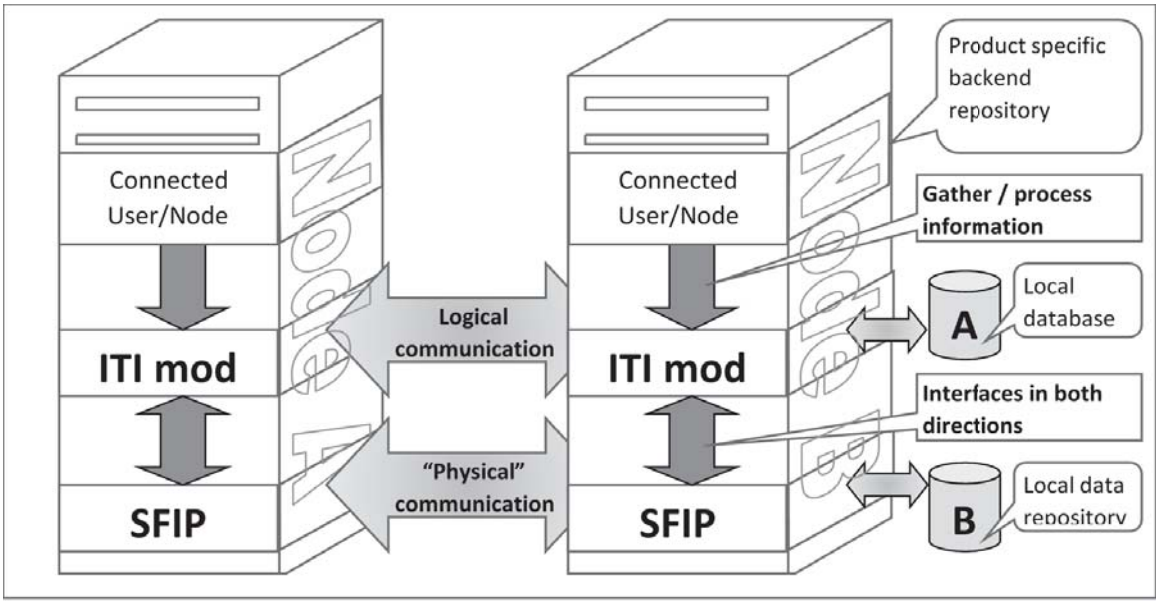


Figure 2: Conceptual architecture (original in Pilgermann et al. 2005)

The exchange of information is comparable with the ISO 7-layer OSI model (Stevens and Wright 1995), (Pilgermann and Blyth 2004). Although, a logical connection is established between the crime-specific modules of the nodes, they are not able to communicate with each other directly. Instead, they are making use of communication facilities provided by the SFIP layer. Furthermore, the crime-specific modules gather and process information from connected analytical workstations. Each node in the overall topology may act as both a source and a consumer of intelligence. However, regarding to roles, certain nodes may only be allowed to either send information or receive information. Each SFIP node maintains its database for storing all information about current and past cases (marked as 'A' in Figure 2). The employed technologies in each layer will be directly addressing the issues of security (both the channel and the content), authentication of nodes and users, and non-repudiation of transactions between the layers and the nodes. Generated logs will be managed as evidence, applying best practice on evidence handling and management as specified by ACPO.

5. Conclusion

Fusing forensic evidence from different cases, performing 'big-data-like' analytical activities, extracting forensic intelligence regarding persons (perpetrators and victims), assets and MOs, is believed to be the future of

resolving computer-crime related activities. Law enforcement strategic vision certainly reflects the above as currently (first quarter of 2016) there is an invitation for tenders for a cybersecurity digital service infrastructure. The aim of the European call is to launch a core service platform that will serve national and/or governmental CSIRTs and CERT-EU. The harmonisation of collaboration procedures between CSIRTs is indeed envisaged to improve cooperation between them and equip them for a better handling of threats to cyber resilience in the European Union.

The proposed system directly addresses the identified requirements in the invitation for tenders. The authors are currently setting up a test-bed that will allow for the prototype development of the proposed system and of the crime-specific analytical modules. The authors are also establishing user groups for the creation and sharing of datasets that will be used for evaluation purposes. When the test-bed goes operational, statistics on the turnaround time of analytical tasks will be generated in order to understand if the proposed system can indeed provide LEAs with an immediate solution to the shortfalls of their current practice. At a later stage, a second data-set on the identified forensic intelligence will be created in order to understand and appreciate the conversion of intelligence enriched cases to successful prosecutions.

References

Angelopoulou, O., Thomas, P., Xynos, K., Tryfonas, T., (2007) Online ID-Theft techniques, investigation and response, International Journal of Electronic Security and Digital Forensics, Volume 1, Issue 1, pp 76-88

BBC, (2015) Number of identity theft victims rises by a third, http://www.bbc.co.uk/news/uk-32890979

Birke, J. (1989) Scientific scene linking, Journal of the Forensic Science Society, volume 29, Issue 4, pp 271-284

CIFAS, (2015) Fraudscape UK fraud trends, United Kingdom, http://www.cifas.org.uk/secure/contentPORT/uploads/documents/External%20-%20Fraudscape%20main%20report%20for%20website.pdf

Chowdhury, T., and Vidalis, S. (2013) Proactively defending computing infrastructures through the implementation of live forensics and website capture in corporate network security. Third International Conference on Cybercrime, Security and Digital Forensics, Cardiff University, Cardiff, UK, June.

Experian, (2015) One in six adults has fallen victim to cyber-crime, http://www.experian.co.uk/blogs/latest-thinking/one-six-adults-fallen-victim-cyber-crime/

Hale, C. (2002) “Cybercrime: Facts & figures concerning this global dilemma”, Crime and Justice International, Volume 18, Issue 65

Harrell, E., (2015) *Victims of Identity Theft, 2014*, U.S. Department of Justice, http://www.bjs.gov/content/pub/pdf/vit14.pdf

ITRC, (2015) Identity Theft Resource Center Breach Report Hits Record High in 2014, *Identity Theft Resource Center*, http://www.idtheftcenter.org/ITRC-Surveys-Studies/2014databreaches.html

Legrand, T., and Vogel, L., (2012) Forensic Intelligence, https://www.academia.edu/1519407/Forensic_Intelligence

Mkomo, (2015) http://www.mkomo.com/cost-per-gigabyte-update)

National Crime Agency, (2015) Forensic Intelligence, http://www.nationalcrimeagency.gov.uk/crime-threats/drugs/forensic-intelligence

Pilgermann, M., Vidalis, S., Blyth, A., (2005) "Inter-Organisational Intrusion Detection Using Knowledge Grid Technology", Journal of Information Management and Computer Security, Volume 14 Number 4.

Pilgermann, M. and Blyth, A., (2004). “Anonymizing Data in a Peer-To-Peer based Distributed Intrusion Detection System - A possible Approach” European Conference on Information Warfare (ECIW), London.

Ribaux, O,. Girod, A., Walsh, S,J,. Margot, P,. Mizrahi, S,. Clivaz, V,. (2003) “Forensic intelligence and crime analysis”, Law Probability and Risk, Volume 2, issue 1, pp 47-60.

Ribaux, O., Margot, P., (1999) “Inference structures for crime analysis and intelligence using forensic science data: the example of burglary”, Forensic Science International, Volume 100, Issue 3, pp 193-210

Statistic Brain, (2015) http://www.statisticbrain.com/average-cost-of-hard-drive-storage/)

Stevens, W. R. and Wright G.R., (1995) TCP/IP Illustrated, Volume 2 - The Implementation. USA, Addison-Wesley Publishing Company

The Register, (2014) http://www.theregister.co.uk/2014/03/25/google_price_slash/

U.S. Department of Defense. (2012) Information operations. Joint Publication: 3-13. Retrieved from http://www.Dtic.Mil/Doctrine/New_Pubs/Jp3_13.Pdf

Law Enforcement Access to Evidence Stored Abroad in the Cloud

Murdoch Watney
University of Johannesburg, Gauteng, South Africa
mwatney@uj.ac.za

Abstract: The legal question arises whether a law enforcement agency may compel a service provider within its jurisdiction to hand over data that provides evidence of the commission of a crime where such data is stored on a foreign server in a country other than the country seeking the data. Many users store their personal information in the cloud for easy and convenient access anywhere and at any time. A law enforcement agency cannot directly access the electronic evidence, but requires the assistance of an intermediary such as the service provider and/or cloud service provider for access to the stored evidence. It has to be established whether there is a legal obligation on a company (cloud service provider) to gather the evidence of a cloud service user stored on a server outside the country's territorial borders on behalf of the law enforcement agency. This is also the focus point in Microsoft Corporation v United States of America (referred to as the "Microsoft-Ireland" test case). If a law enforcement agency must use Mutual Legal Assistance Treaties (MLATs) for international law enforcement, it should be ascertained whether the MLAT process is fast and cost effective. If the MLAT process is not efficient, a country may employ data localisation as an alternative law enforcement method which will impact on cloud computing. It is relevant to establish which interests weigh the most: the safety and national security of the requesting country or the privacy rights in protecting the personal information of the user or the sovereignty of the country where the evidence is stored? Although the focus point under discussion may not be a new concern, it has become a very relevant and contentious issue that necessitates legal clarification as the Internet is not only a global network but users increasingly store a lot of personal information in the cloud.

Keywords: law enforcement, stored electronic evidence, international law, cloud computing, data localisation laws, mutual legal assistance treaties

1. Introduction

The legal question arises whether a country's law enforcement agency may extraterritorially access data stored in the cloud in another country.

Cloud computing is the result of technological development and globalisation. In its simplest terms it is about storing and retrieving personal data from the Internet (Griffith 2015). The process entails the storage of data in multiple locations to reduce costs and increase speed as the data may be accessed over the Internet anywhere and anytime, as the data is not stored on the local hard drive of the user's computer (Griffith, 2015). Cloud service providers such as the US company, Microsoft has for example data centres in various countries to ensure that data is close to customers for quick and smooth access (Segal, 2015).

The EU recognised the importance of cloud computing by adopting a strategy in 2012 for "Unleashing the Potential of Cloud Computing in Europe". The strategy aims to unify rules and standards related to cloud computing within Europe and to promote and facilitate faster adoption of cloud computing throughout all sectors of the economy. The relevance of the issue at hand was highlighted in a study conducted in 2015 by the Centre for European Policy Studies on the challenges facing EU rule of law and fundamental rights regarding access to electronic data by third country law enforcement authorities (Carrera et al, 2015). In the study the emphasis was on EU-US relations in respect of mutual legal assistance and evidence gathering for law enforcement purposes (Carrera et al, 2015).

This is also the focus point in Microsoft Corporation v United States of America (referred to as the "Microsoft-Ireland" case) which will be discussed hereafter as note should be taken of the arguments for and against remote access to stored evidence on a foreign server. The outcome of this case will have a global impact on the extent of law enforcement powers and has therefore evoked a lot of interest in countries and customers outside the United States (US) and Ireland which are the main role-players in the matter. It may also have legal ramifications for cloud computing. Although the discussion focuses on the primary legal question, it also touches briefly on various inter-related legal issues such as:

- Sovereignty of the country which stores the information;
- Territorial jurisdiction;
- Extraterritorial jurisdiction;

- Nationality of a customer;
- Privacy protection of the personal information (also referred to as informational privacy or data protection) of a customer;
- Ownership of data stored in the cloud;
- Localisation of data laws; and
- Applicability of mutual legal assistance by means of Mutual Legal Assistance Treaties (MLATs) or the European Investigation Order (EIO).

2. Conceptualization of terminology

Conceptualization of the relevant terminology provides a background to the discussion. The main issue will be deliberated within the context of the international law and not from the perspective of the domestic US procedural law. Where relevant however, reference will be made to the US domestic law such as the Stored Communications Act (SCA) and whether a warrant issued in terms of domestic legislation is applicable outside its borders. The issue of whether data stored outside a country's borders may be seized for law enforcement purposes will be debated with reference to issues such as sovereignty, territorial jurisdiction and extraterritorial jurisdiction.

International law may be defined as a body of rules and principles which are binding upon states in their relations with one another (Dugard, 2010). Sovereignty empowers a state to exercise the functions of the state within a particular territory to the exclusion of other states (Dugard, 2010). Jurisdiction is an important aspect of sovereignty which defines these functions (Dugard, 2010). In general states are confined to the exercise of their functions within their own territories, but with cross-border crime and the gathering of evidence, states may have an interest in extending their jurisdiction beyond their territorial limits to cover persons, property and evidence in other countries (Dugard, 2010).

The continuous growth in the volume of cross-border communication increases the need for law enforcement agencies to seek access to electronic evidence across national borders (Swire and Hemmings, 2015). It is therefore important to define cross-border law enforcement powers in accessing stored evidence in the cloud.

3. Microsoft-Ireland case as a test case

3.1 Facts of the case

In December 2013 the US Department of Justice (DOJ) served the US based Microsoft with a warrant requiring it to hand over the emails of a Microsoft customer suspected of drug trafficking. The warrant was issued in terms of the Stored Communications Act (SCA) which was enacted as part of the 1986 Electronic Communications Privacy Act (ECPA) (Ely, 2015).

Microsoft refused to turn over the emails on the basis that they were stored on servers at a data centre in Ireland and argued that the warrant did not have extraterritorial application.

The reason why the data is stored outside the US appears to be a matter of network design. Microsoft has designed its network to maintain the emails of individuals who signed up for accounts using foreign country codes on servers in Ireland instead of the US. When the target of this investigation set up a Microsoft e-mail account, he entered a country code outside the US which led Microsoft to store the data in Ireland (Kerr, 2014). It is not clear whether the suspect is a US national or a foreigner although it has been surmised that the suspect is a non-US national and most probably an Irish citizen residing in Ireland (MacCarthy, 2015). As will be illustrated hereafter, the nationality of the customer is relevant.

Microsoft was of the opinion that the data was protected by the laws of the country where its servers are located and argued that the DOJ should work with the Irish authorities to obtain access to the data. Microsoft placed the emphasis on the location of the data to determine territoriality. In view of this argument the request of the DOJ to retrieve the evidence would violate the sovereignty of Ireland as compliance with the warrant would result in the extraterritorial application of the warrant.

Microsoft also argued that the evidence requested by the DOJ consisted of emails which represent the content of personal communications and could not be considered as business records such as a record of banking transactions, a hotel bill or a list of phone numbers (Segal, 2015). Microsoft contended that it did not own the emails but that the ownership of the emails resided with the email user (Thielman, 2015). Accessing the emails would therefore seriously violate the privacy rights of the customer.

Microsoft's opinion should also be seen against the background of Snowden's 2013 revelations of US surveillance practices which had an unfortunate spin-off that is still being felt today (Watney, 2015). Kerr (2014) makes a valid observation that in a post-Snowden age non-US Internet users do not trust the US legal system with the result that US providers and in this instance, the cloud service provider, want to keep foreign customers on board by promising them that the process of the law pertaining to foreign jurisdiction would be adhered to.

Microsoft argued that the DOJ should obtain the information through the Mutual Legal Assistance Treaties (MLATs) which provide bilateral frameworks for law enforcement co-operation. In this instance the DOJ would have to operate in terms of the Mutual Legal Assistance Treaty with Ireland and obtain a court order from an Irish judge in order to obtain access to the emails.

The DOJ argued that as long as a US-based company exercises custody and control over evidence which is the subject of a court-issued warrant, the physical location where the information is stored and the nationality of any person associated with information are both immaterial (Ely, 2015). The warrant issued in terms of the Stored Communications Act encompassed the serving of a warrant in the US and searching premises in the US (Thielman, 2015). Territoriality is therefore determined by the service provider's location and not the data location. This would imply that Microsoft may unilaterally and remotely retrieve the evidence relating to a crime committed on US territory (MacCarthy, 2015).

The DOJ also contended that the emails should be treated as business records of the hosting company and as a result only a search warrant would be needed in order to compel the provision of access to it no matter where they are stored.

Although the question within the US national legal framework is whether the Stored Communications Act and Fourth Amendment to the US Constitution apply to data stored overseas, the question should also be seen in a broader context from the perspective of the international law. Central to the case is the debate on whether granting remote access to stored data on a foreign server for law enforcement purposes will be harmful to the future of the Internet, privacy, respect for borders and public safety (Segal, 2015).

3.2 Argument in favour of remote access to stored data on a foreign server

Woods (2015) is of the opinion that granting access to stored data on a foreign server does not violate the sovereignty of a country. The DOJ is merely requesting a company doing business in the US to produce evidence stored offshore as it has done in dozens of offshore banking cases. It appears that Woods supports the argument that the provider's location is determinative and not the location of data. Such a request must however be subject to the requirement of a warrant being issued by an objective magistrate (Woods, 2015).

Granting remote access to stored data on a foreign server may be more advantageous to the future of the Internet than denying access. If a warrant cannot be applied extraterritorially to access data in another country, it may result in the adoption of data localization laws. Data localisation laws will be discussed at paragraph 6 hereafter.

3.3 Argument against remote access to stored data on a foreign server

The consequence of remote access to evidence is that it may bypasses existing legally-binding channels which may result in legal uncertainty and mistrust in transatlantic relations, private sector-public institution relations and the public (Carrera et al, 2015). Furthermore, if the US government is allowed access to the data in these circumstances, then other countries will follow suit. A country's law enforcement agency may request a service provider with a presence in its country to hand over data stored abroad and the request may not be dependent on a warrant requirement as is the case in the US (Segal, 2015).

If the US law enforcement agency may access the data stored by US providers abroad, irrespective of the nationality of the customer, then customers will be encouraged to encrypt their communications and/or US cloud operators may store the data in a non-US data centre (Wang, 2015). Snowden's revelations inter alia resulted in cloud service users re-assessing the risks of storing data in the cloud (Kronqvist and Lehto, 2015).

Cognisance should be taken of article 32 of the 2001 Budapest Cybercrime Convention, a multilateral agreement, which provides that a law enforcement agency may access data extraterritorially if the data is publicly available (open source) or with the lawful and voluntary consent of a person legally entitled to disclose the data. The Cybercrime Convention Committee (T-CY) adopted Guidance Notes in 2014 regarding the application of trans-border access to data with reference to article 32. The Guidance Notes provides that service providers are unlikely to be able to consent validly and voluntarily to the disclosure of their users' data under article 32. In terms of the Guidance Notes service providers will only be holders of such data and they will not control or own the data and would therefore not be in a position to consent. The best solution for obtaining cross-border evidence would be through Mutual Legal Assistance Treaties which involve all countries on a global level.

4. Privacy of personal information and security

At the core of law enforcement powers are privacy and security and therefore a discussion on access to stored data in the cloud will be incomplete without having regard to privacy protection of personal information (also referred to as information privacy or data protection). Heyink (2015) correctly states that privacy today is the "most burning jurisprudential issue globally and pervades the political, economic, societal and technological landscape, shaping approaches to existing and new law in the information society at every turn."

Individuals disclose daily large amounts of personal information on the Internet as well as to private companies whom collect and process massive quantities of electronic data. It may potentially be possible for a country to access personal information which has been captured and processed in another country if there are not effective and enforceable safeguards in place to ensure privacy protection.

The European Union has taken the lead in protecting the privacy rights in respect of personal information. Many countries look at the EU for guidance in respect of the protection of personal information. There are many dissimilarities between the EU and the US regarding data protection (Carrera et al, 2015). It appears that the US lean more towards the right to access of information than the protection of privacy (De Sadler and Esselaar, 2015).

The EU commitment to data protection as a core value was confirmed on 15 December 2015 when the European Commission introduced reform to data protection which consists of two laws, namely the General Data Protection Regulation (GDPR) and the Data Protection Directive for the Police and Criminal Justice Authorities. The GDPR which replaces Directive 95/46/EC provides extensive privacy protection to personal information and makes specific provision for cloud service providers. The Data Protection Directive for the Police and Criminal Justice Authorities ensures a legal framework for the sharing of information by means of the transfer and processing of data for law enforcement purposes while protecting the privacy rights in personal information. The need of a proportionate approach to the processing and sharing of data should not override the EU fundamental rights to privacy and the protection of personal data (Bowman, 2015).

In terms of the GDPR a single law and a single data protection authority (DPA) will be applicable to all 28 EU member countries and a company would only be answerable to the data supervisory authority in the country where they have their main establishment. Individuals will have the right to refer all cases to their home national data protection supervisory authority where their personal data is processed outside their home country. A company based outside the EU offering services in the EU must comply with EU data protection laws. The latter was already reflected in a European Court of Justice (ECJ) judgement in early October 2015 where the court ruled that if a company operates a service in a country it can be held accountable by that country's national data protection agency despite not being headquartered there (Lomas, 2015). Requesting a US company to retrieve the emails of a European national stored in Europe could be seen as a disregard for EU privacy protection laws (Arthur, 2015).

Taking into consideration the strict data protection rules, obligations and safeguards, it would not be easy for a cloud service provider to move the data outside the EU although the GDPR does not prohibit the transfer outside

the EU. The cautious approach of the EU to data transfer was illustrated when the ECJ in October 2015 declared the "safe harbour" data transfer agreement between the US and EU invalid as the US did not afford adequate levels of protection of personal data (Lomas, 2015). The "safe harbour" agreement between the US and the EU meant that a US company who processed a European citizen's personal data was under the same protection as if it were still located in Europe on a European-owned system.

Although the data protection reform was discussed long before the 2013 Snowden revelations, the revelations heightened privacy awareness and the necessity of implementing an effective and enforceable legal framework in protecting personal information. Unfortunately, the tension between the demands of law enforcement authorities to access information to address the increase in cybercrime on the one hand and the protection of personal information of the user on the other hand will not subside. The service provider cannot be seen as a mere extension of the state apparatus (law enforcement agency) for purposes of access to communications for law enforcement purposes. The service provider must still access whether a request of the state meets the requirements of the right to privacy. If the service provider has a strong suspicion that it does not, it cannot simply disobey, but have a duty to test their view in court (Bilchitz, 2016). Bilchitz (2016) is of the opinion that such a duty to exercise independent judgment will result in publicity and transparency over requests by the government for personal information.

It is clear that achieving a balance between security and privacy and free speech will have to be addressed within a legal framework aimed at legal certainty and trust and in this regard, mutual legal assistance between countries in accessing electronic evidence for law enforcement purposes are of vital importance.

5. Mutual Legal Assistance Treaties (MLATs)

The internationalisation of crime has made national law enforcement agencies increasingly dependent on international co-operation. This has resulted in states entering into both bilateral and multilateral mutual assistance treaties (Dugard, 2011).

Carrera et al (2015) define mutual legal assistance (MLA) as a "classical treaty-based mechanism allowing for foreign law enforcement cooperation and assistance in ongoing criminal investigations and proceedings, while respecting the notions of jurisdiction and national sovereignty in criminal matters".

The importance of MLATs is reflected in the European Investigation Order (EIO) which regulates the exchange of evidence between EU members (Walden, 2013). The Directive on the European Investigation Order (EIO) will become from 22 May 2017 the sole legal instrument regulating the exchange of evidence and mutual legal assistance between EU member states (Carrera et al, 2015).

There also exists a 2003 EU-US MLA but the EU-US MLA is subject to the EU data protection legal framework.

Some commentators (Walden, 2013; Kerr, 2014; Segal, 2015) are of the opinion that the processes involved with the application of mutual legal assistance treaties (MLATs) are time-consuming, complex and not effective as the data could easily be moved to another country. This poses a serious challenge under circumstances when obtaining electronic evidence speedily is critical to criminal investigations (Kerr, 2014; Segal, 2015).

Carrera et al (2015) are of the opinion that the practical obstacles facing the implementation of mutual legal agreements may be overcome through a combined approach focused on bilateral case consultations, day-to-day contacts, stronger political commitments, more effective use of existing tools and sound financial, technological and human resource investments in their implementation.

Swire and Hemmings (2015) indicate that the MLAT process has become important for a global and inter-operable Internet against calls for data localization and other stricter national controls of the Internet. They emphasize the importance of an effective and speedy MLAT and make various suggestions in improving the MLAT processes while at the same time ensuring the protection of privacy and free speech. The authors emphasize that the present application of MLAT require urgent attention if it is to remain a primary means of providing trans-border access to data for law enforcement (Walden, 2013; Swire and Hemmings, 2015). An alternative to MLATs may be data localisation to ensure that a country's law enforcement authority has access to evidence.

6. Data localisation laws

Data localisation laws require mandatory storage of data on servers physically located within the borders of a country. The motivation for such data localisation laws is based on the protection of security and privacy. It may also be attributed to Snowden's revelations as a measure to prevent unauthorised access to data by another country (Dhont and Woodcock, 2015).

Russia implemented data localisation laws which became effective on 1 September 2015. All legal entities must store and process the personal data of Russian citizens on servers located within the Russian territory (Bauer, 2015). Personal data can be transferred as long as the primary database used for collection, storage and processing remains in or will be transferred to Russia (Bauer, 2015). Other countries such as South Korea, Brazil, Vietnam and Indonesia already have localisation laws in place (Dhont and Woodcock, 2015).

Data localisation laws may hinder or restrict the transfer of data across national borders (Dhont and Woodcock, 2015). In this regard, the GDPR provide stringent requirements for transfer of personal information outside the EU borders and the laws of the EU may in effect act as data localisation laws (Bauer et al, 2014; Dhont and Woodcock, 2015).

Data location laws may be advantageous to security and privacy but it will impact negatively on the economy. For example, a number of administrative regulatory barriers could be introduced through additional legal obligations that increase compliance costs, such as stricter consent requirements, a right to review personal information held by firms, the requirement to notify a market regulator and/or data subject regarding potential security breaches (Bauer et al, 2014). Some measures are institutional such as the requirement to appoint a data privacy officer within the organisation while others increase business risks by introducing sanctions for non-compliance (Bauer et al, 2014). A company and in this instance a cloud service provider will be subject to location demands in every jurisdiction it operates which would have enormous cost implications (Woods, 2015). Companies will also not have the flexibility to store information wherever network efficiency dictates which is one of the characteristics of cloud computing. The consequence of data localisation may be a fragmented Internet along national borders (Bauer et al, 2014).

Governments will have to decide whether data localisation laws are to the advantage of the growth and accessibility of the Internet.

7. Recommendations

There is a need to address international rules and processes and in particular the MLATs.

On domestic level countries will have to address the challenge by means of new legislation or updating existing laws to keep pace with the developments brought about by the Internet age.

When the US Electronic Communications Privacy Act was enacted in 1986, storing large amounts of data on a remote server by means of cloud computing did not exist. Authors such as Kerr (2014) and Segal (2015) are of the opinion that the US congress should provide an outline of the application of the Stored Communications Act abroad. The proposed US Law Enforcement Access to Data Stored Abroad (LEADS) Act provides that where a US national uses a US service provider, the DOJ may access the data by means of a warrant irrespective of where it is stored. The US warrant pertaining to a US national would therefore have extraterritorial application. However, if a non-US national stores his emails abroad through a provider who have offices in the US, the data may only be assessed by means of MLATs (Kerr, 2014). The LEADS Act disregards the location of data in respect of a US citizen but makes it determinative when dealing with a non-US citizen. If information of a non-US citizen is stored in the US, then it may be subject to a US warrant. MacCarthy (2015) suggests that jurisdiction rules focus on a user citizenship or location rather than where the data is stored. Kerr (2014) also states that nationality should play a role.

In my opinion a government's powers should be limited to its territory and/or nationals within its territory. The US LEADS Act provides that a government may access the data of its own nationals stored abroad and therefore the cloud has nationality. However, it may also be argued that in instances where a national's communications are outside the borders of the country, the rules and principles of the international law should be applied.

8. Conclusion

Abuse of the Internet for cyber-dependant and cyber-enabler crimes cannot be tolerated. The gathering of electronic evidence as mechanism to prove the commission of a crime is an essential tool in the investigation of criminal activities. Any delay in speedy cross-border assistance may hamper the gathering of electronic evidence with the consequence that the abuse of the Internet may proliferate to the detriment of its future.

Irrespective of the outcome of the Microsoft-Ireland case, access to data stored abroad and specifically in the cloud must be addressed on a domestic and international level. It is clear that several issues require clarification. Governments face many conflicting interests namely the privacy protection of personal information (data protection) on the one hand but at the same time ensuring the security of citizens. Countries may move in the direction of data localization laws to give effective protection to personal information and to circumvent cross-border challenges in addressing law enforcement but the development of a fragmented Internet across national borders would have to be weighed against the economic and/or social growth of the Internet in general.

References

Arthur, C. (2015) "Safe harbour ruling illustrates growing chasm between US and EU", [online], http://www.theguardian.com/technology/2015/oct/06/safe-harbour-ruling-growing-ch.

Bauer, M. (2015) "EU-US Safe Harbour and forced data localization: lessons from Russia", [online], http://wwwl.euractiv.com/sections/digital/eu-us-safe-harvour-and-forced-data-localisation-lessons-russia-318606.

Bauer et al. (2014) "The costs of data localisation: friendly fire on economic recovery", [online], www.ecipe.org/app/uploads/2014/12/OCC32014_1.pdf.

Bilchitz, D. (2015) "Privacy, surveillance and the duties of corporations", *Journal of South African Law,* pp 45 – 67.

Bowman, J. (2015) "Paris attack bring police directive negotiations back into spotlight", [online], http://iapp.org/news/a/paris-attacks-bring-police-directive-negotiations-back-into-spotlight/.

Carrera, S. et al. (2015) "Access to electronic Data by Third-Country Law Enforcement Authorities: Challenges to EU Rule of Law and Fundamental Rights", [online], http://www.ceps.eu.

De Sadler, E. and Esselaar, P. (2015) *Protection of Personal Information Act*, Juta and Co. Ltd, Cape Town, p 3.

Dhont, J. and Woodcock, K. (2015) "Data localization requirements: Growing trends and impact for company compliance", [online], http://www.corporatecompliance.org.

Dugard, J. (2011) *International Law: A South African Perspective*, Juta and Co. Ltd, Cape Town, pp 1, 146, 148 – 149, 237 - 239.

Ely, A. (2015) "Second Circuit Oral Argument in the Microsoft-Ireland Case: an Overview", [online], https://www.lawfareblog.co/seond-circuit-oral-argument-microsoft-ireland-case-ov.

Griffith, E. (2015) "What is cloud computing?", [online], http://cpf.cleanprint.net/cpf?action=print&type=filePrint&key=pcmag&url=http%.

Heyink, M. (2015) "Why are South African lawyers remaining in the dark with POPI?" *De Rebus*, August, pp 31 – 33.

Kerr, O. (2014) "What legal protections apply to e-mail stored outside the U.S?", [online], https://www.washingtonpost.com/news/volokh-conspiracy/wp/2014/07/07/what-legal.

Kronqvist, J. and Lehto, M. (2015) "Adopting encryption to protect Confidential data in Public Clouds: a review of solutions, implementation, challenges and alternatives", in Abouzakhar, N *Proceedings of the 14th European Conference on Cyber Warfare and Security*, Academic Conferences and Publishing International Limited, Reading, UK, pp 151 - 158.

Lomas, N. (2015) "Europe's top court strikes down 'safe harbor' data-transfer agreement with US", [online], http://techcrunch.com/2015/10/06/europes-top-court-strikes-down-safe-harbor-data-transfer-agreement-with-u-s/

MacCarthy, M. (2015) "A better way to think About the Microsoft-Ireland Case", [online], http://blog.siia.net/index/php/2015/09/a-better-way-to-think-about-the-microsoft-irela.

Segal, A. (2015) "Does a U.S. warrant apply to data stored on a Foreign Server?", [online], http://www.newsweek.com/does-us-warrant-apply-data-stored-foreign-server-3.

Swire, P and Hemmings, J.D. (2015) "Re-engineering the Mutual Legal Assistance Treaty Process", [online], http://www.heinz.cmu.edu/~acquiti/SHB2015/Swire.docx.

Thielman, S. (2015) "Nationality in the cloud: US clashes with Microsoft over seizing data from abroad", [online], http://www.theguardian.com/us-news/2015/sept/02/microsoft-us-government-cloud-co

Wang, C. (2015) "Data sovereignty: a layman's guide to the Microsoft Ireland Case", [online], http://www.cipercloud.com/bog/data-sovereignty-laymans-guide-microsoft-ireland-.

Watney, M.M. (2015) "The Legal Conundrum Facing ISPs in Social Media Policing Against Extremism", in Abouzakhar, N. *Proceedings of the 14th European Conference on Cyber Warfare and Security*, Academic Conferences and Publishing International Limited, Reading, UK, pp 300 – 306.

Walden, I. (2013) "Law Enforcement Access to Data in Clouds" in Millard, *C Cloud Computing Law*, Oxford University Press, New York, USA, pp 285 – 310.

Woods, A. (2015) "Lowering the temperature on the Microsoft-Ireland case", [online], http://www.brookings.edu/blogs/techtank/posts/2015.0921-lowering-temperature-microsoft-case.

Clandestine Cell Based Honeypot Networks

Cagatay Yucel, Ahmet Koltuksuz and Huseyin Yagci
Department of Computer Engineering, Yaşar University, Izmir, Turkey
cagatay.yucel@yasar.edu.tr
ahmet.koltuksuz@yasar.edu.tr
huseyin.yagci@stu.yasar.edu.tr

Abstract: A Clandestine Cell is a type of an intelligence organization where a cell only knows the immediate superior and the associated members of itself. This kind of organizational structure is used by intelligence agencies throughout the world to provide security against a breach, thus ensuring the safety of the members. This well-known intelligence organization is applied to solve an advanced cyber security issue. A relatively new kind of a cyber threat known as an Advanced Persistent Threat (APTs) has been around for some time now, Stuxnet being the very first identified. There are several points to consider when identifying the characteristics of an APT, such as the aim, its interactions with Internet, way of collecting information, operations they do disrupt and concealment mechanisms utilized. An important aspect is whether it is statistically analyzable or dynamically identifiable, that its communication patterns need to be inspected to identify the characteristics. The traces of an APT might be identified this way. In this research, a honeypot network with a communication policy based on a clandestine cell is introduced. Each honeypot only knows a hub. And a hub only knows the main malware analysis server. By utilizing this approach, the communications are hidden from possible attackers without compromising the main server. In each honeypot server, dead-ends are created and implemented in the honeypot servers. Advantages and ramifications are discussed regarding the types of malware. It is aimed to create yet another taxonomy of malware regarding the network activities as they are being trapped by our introduced honeypot network. A clandestine cell format is one of its kind within organizations. This is the very first time that such kind of format is being applied to honeypot design for APT hunting. This is the paper in which an intelligence organizational structure meets with a network architecture in order to solve a very hard to crack cyber security problem. The idea itself is a new and untried one.

Keywords: clandestine cell, honeypots, advanced persistent threats, clandestine network organizations

1. Introduction

Malicious activities of computer systems started almost at the same time with the invention of *Von Neumann* computer architecture. In this computational model, John Von Neumann foresaw a program that is able to self-reproduce on the memory, which is considered the first computer virus (Neumann, 1969). On 1988, a harmful computer program that was able to self-reproduce on networks named *Morris Worm* effecting a large number of computers have been unleashed (Eisenberg et al., 1989). From then on, advances have been made on both sides of this war: the security professionals and the attackers.

A relatively new kind of cyber threat, Advanced Persistent Threats (APTs) has been around for almost 8 years now (Langner, 2011). Mainly targeting the industrial control systems, political institutions and critical infrastructures, this type of threats are way ahead of conventional defense mechanisms. They are advanced as in they use vulnerabilities which have not been identified and combining social engineering techniques with computer intrusion technologies. They stay under the radar until they reach their targets (Saud & Islam, 2015).

These advanced threats require advance proactive defense techniques to cope with them. One of the recent solutions of such is Honeypots. Honeypots are information systems used for exploiting the attacker by luring them with decoys. A honeypot is expected to be probed, attacked and exploited (Spitzner, 2003). A honeypot network or a honeynet is a collection of honeypots for large networks, collecting information about most recently developed attacks as well as the attackers. One of the biggest challenges whilst collecting information is that the network should not be compromised.

In this research, a well-known intelligence organization, Clandestine Cell Network is implemented on honeypot networks in order to provide the maximum secrecy of the honeypot network when a cell or a single honeypot is compromised. This paper is organized as follows: Section 2 presents the characteristics of APTs, Section 3 describes and defines honeypots. In Section 4, proposed model is explained with the organizational charts and Section 5 addresses the advantages and ramifications of such system and concludes the paper.

2. Advanced persistent threats (APTs)

An advanced persistent threat is an adversary that utilizes advanced levels of expertise, significant resources and objective specific tools to execute its objectives by series of attacks. These attacks may include cyber, physical and deceptive techniques. These objectives typically include gaining access to the targeted infrastructure for the aims of gathering, disrupting or modifying the critical aspects of a mission, program, or organization or infiltrating the infrastructure to accomplish its objectives in the future.

An advanced persistent threat must have the following characteristics:

- It should try to achieve its objectives repeatedly over an extended period of time,
- It should overcome the defending mechanisms,
- It should accomplish the necessary infiltration and maintain a connection with the Command and Control.

2.1 Characteristics of known APTs

Stuxnet

Stuxnet is the first known megahit APT attack on the Industrial Control Systems (ICSs). The APT designed and developed for disrupting Iran's nuclear enrichment program. The APT directly affected programmable logic controllers (PLCs) and caused overloading on the centrifuge that is used in the enrichment operation. There were many suspicious dead-end IP addresses to hide the source code of Stuxnet. Waves of attacks started from 2009 and continued to 2010 (Falliere, Murchu, & Chien, 2011).

Operation Aurora

Operation Aurora is an attack which shows APT properties. Victims are infected with social engineering techniques and a malware known as **Trojan.Hydraq** is downloaded via a zero-day exploit on the web browser Microsoft Internet Explorer. Finally, the malware created a backdoor on the infected computer and gave access to the sensitive data. Operation Aurora was publicly discovered in 2010 (Varma, 2010).

GhostNet

GhostNet is an APT attack that involves a malicious network and a malware. The malware forces infected nodes to send an email with an attached Trojan named **"gh0st RAT"** and exploitation code to the victims on the network as a crafted email. Moreover, "gh0st RAT" gains root privileges on the host computer. Starting from 2007, the incident had effected more than 1,300 computers in 103 countries. There were also military, diplomatic and political networks known to be infected (Information Warfare Monitor, 2009).

Taidoor

Taidoor is a Trojan that has been used since 2008. Taidoor's victims are government agencies, corporate entities, and think tanks, especially those with interests in Taiwan and US (Doherty & Krysiuk, 2011). Taidoor has been spreading itself as an email attachment and once the email is opened the backdoor is injected into the memory as an operating system service and connection is established with the Command and Control server of the Trojan.

IXESHE

IXESHE is an APT that is notable for targeting East Asian governments, electronics manufacturers, and a telecommunications company in 2011. Adobe Acrobat Reader, and Flash Player, Microsoft Excel exploits; CVE-2009-43243, CVE-2009-09274, CVE-2011-06095, CVE-2011-06116, CVE-2009- 43247, CVE-2011-06098, CVE-2009-3129 are used after infection (Sancho, Torre, Bakuei, Villeneuve, & McArdle, 2012).

Poison Ivy (PIVY), "Nitro"

Nitro is a cyber-incident that is targeted to the chemical industry and government agencies in 2011. Poison Ivy is a totally free windows based remote access tool. The tool and Internet Explorer based zero day exploit are used in this incident. Adversaries generate code for maintenance and networking with using PIVY tool and the code is injected into the running instance of an Internet Explorer process. Based on adversaries' configuration, a remote shell is activated over TCP ports (Fireeye, 2014).

Duqu

Duqu is announced as the latest discovered version of Stuxnet. The reason that this relation between Duqu and Stuxnet has been made is due to Duqu's aim to collect valuable data about industrial infrastructure. This incident doesn't have any kind of remote access Trojan (RAT) to control or effect the industrial system. It focuses on stealing valuable data. The malware uses Microsoft Word files which contains a zero day (CVE-2011-3402) on targets (Symantec, 2011).

Flame

Flame is an APT mainly designed for espionage activities. It has targeted Microsoft Windows OS computers and been stealing critical information by utilizing key logging, capturing screen shots and switching microphone and camera on to record some valuable information. It also, searches for available neighbor computer and turns the first infected computer into a proxy server for Windows Update [6]. Adversaries used the same zero day vulnerabilities as Stuxnet which are Print Spooler (MS10-061) and Windows Shell (MS10-046) (Bencsáth, Pék, Buttyán, & Félegyházi, 2012).

Red October

Red October is a wide scaled cyber espionage operation that has been effective in more than 39 countries. It is discovered in 2012 by Kaspersky Labs. Focused targets are critical infrastructures such as governmental, military, energy information systems. The aim of this incident is gathering assets. Spear phishing techniques are used alongside with the malware embedded in email attachments. Encrypted servers are used in C&C stage (Chavez, Kranich, & Casella, 2015).

MiniDuke

MiniDuke is an information stealer type of malware that has effected more than 23 countries. MiniDuke uses malicious crafted emails for spreading and it has a unique communication mechanism with the Command and Control servers via encrypted URL on twitter. If twitter accounts are blocked and unreachable, it uses Google Search to reach its C&C servers (Virvilis, Gritzalis, & Apostolopoulos, 2013).

In Table 1 below, aforementioned APTs are summarized. Zero day exploits are given in the numbering form of *Common Vulnerabilities and Exposures* dictionary ("Common Vulnerabilities and Exposures," 1999).

Table 1: APT characteristics table for designing the honeypot network

Name	Time	Port/s	Protocol	Zero-Day Exploits	Target System(Attack Vectors)	Functionality
GhostNet	2007	80, 8000, 4501,	HTTP	CVE-2006-2492, CVE-2006-2492	Government, Energy industry, Military, Universities	Taking full control
Taidoor	2008	80	HTTP	CVE-2009-1129, CVE-2011-0611, CVE-2011-2100	Multinational big companies, Think Tank,	Information gathering
IXESHE	2009	80, 443, 8080	HTTP, HTTPS	CVE-2009-4324, CVE-2009-0927, CVE-2011-0609, CVE-2011-0611	Multinational	information stealing, Remote code execution
Poison Ivy "Nitro"	2011	3460, 80, 443, 8080,	HTTP, HTTPS,	CVE-2012-0158, CVE-2009-4324, CVE-2013-0422, CVE-2013-1347, CVE-2011-3544	Chemical industry, Government agencies	information stealing, Remote code execution

Name	Time	Port/s	Protocol	Zero-Day Exploits	Target System(Attack Vectors)	Functionality
		1863, 8000				
Duqu	2011	80, 443	HTTP, TCP	CVE-2011-3402	Multinational	information stealing
Flame	2012	80, 443, 22	HTTP, HTTPS, SSH	MS10-061, MS10-046	Middle East	information stealing, Remote code execution
Red October	2012	40080	TCP	CVE-2009-3129, CVE-2010-3333, CVE-2012-0158, CVE-2011-3544	Government, Energy industry, Military, Universities	High level cyber espionage, information stealing
MiniDuke	2013	443, 80, 8080	HTTP, HTTPS	APSB13-08, CVE-2013-0643, CVE-2013-0648,	Multinational	information stealing
Stuxnet	2010	80	HTTP	MS10-061, MS08-067, MS10-062, MS10-046, CVE-2010-2568	Industrial control systems and (PLC)	Disruption
Operation Aurora	2010	80	HTTP	CVE-2010-0249	Multinational big companies	information stealing

3. Honeypots

The term "*honeypot*" or "*honeytrap*" comes from human intelligence (HUMINT) terminology and refers to a strategy where an attractive male or female agent is used to seduce individuals and exploit this sexual relationship to force individuals to cooperate with them. History shows that even a well-trained, most clever and patriotic person can fall into this trap if set properly. The honeypot driven espionage operations have been heavily referenced in intelligence literature (Digby Diehl & Clarridge, 1997; Earley, 1997; Wright & Greengrass, 1988). A honeypot in computer terminology is a decoy based information system designed to lure the attackers into its traps and try to log information about malwares and attackers. The concept of the honeypots and deception in information systems is first introduced by Fred Cohen's Deception Toolkit (Cohen, 1998). The attackers often search for the vulnerabilities in an information system and attack the weakest points, therefore a vulnerable system to attract the attackers shall be used as a honeypot. After an attack is conducted, the aim is to detect an attack, identify the vulnerability and find out the attacker and Command and Control (C&C) center of the attack.

Intrusion Detection Systems (IDS) is a conventional tool that monitors the network traffic and searches for a potential malicious activity. A major shortcoming of IDS is that these systems are equipped with a pattern database of known attacks and thus the pattern matching is done by a technique known as deep package inspection. However, this operation is quite resource consuming and is potentially a victim of Denial of Service (DoS) attacks. Another ramification of the IDS is that it is highly detectable and therefore when an attacker knows that the traffic is monitored and inspected, then he can force IDS to blind its sensors by false negatives and/or by applying network flooding techniques. On the other hand, a honeypot system acts like any other server on the network with an unused address. Therefore when a malicious attempt or a breach on the honeypot server is spotted then it is very high probability that such an attempt is an actual malicious activity (Fanfara, Dufala, & Radušovský, 2013).

Honeypots can be installed in multiple numbers to a network which forms a Honeynet. A Honeynet can be used for wider networks where one honeypot would not suffice (Jasek, Kolarik, & Vymola, 2014). In this case, the communication of honeypots become critical: it must be accomplished in a stealthy way in order to hide the existence of these fake systems and it is as important as is the information of attacks must be disseminated as quickly as possible to alert the overall system which is being protected. This research proposes a communication protocol inspired by intelligence agencies described in the next section.

A honeypot should interact with the attacker in order to trick the attacker to believe that it is a legitimate system. Regarding the level of interaction honeypots can be divided in two:

- Low-interaction honeypots

This type of honeypots only emulate a few steps and replies of the vulnerable network protocol and network stack that is being imitated. It is easy to deploy and use. However it is easy for an attacker to detect one.

- High-interaction honeypots

High-interaction honeypots have a vulnerable operating system and network services fully implemented in it, generally have reduced operating system kernels. It is the most complex type having the ability of collecting all malicious activity with the possibility of takeover by an attacker as a disadvantage. They are generally monitored by an external IDS for such possibility.

A significant advantage of utilizing these systems is the possibility to detect new types of attacks and vulnerabilities that are used by attackers. Honeypots are attracting computer security researchers as they lure the attackers as well, many interesting research have been done on honeypots. A hybrid honeypot system is proposed in (Fanfara et al., 2013) where a low interaction honeypot is combined with a high interaction one in cooperation with an IDS. Hardware abstraction methodology is proposed in (Zhang, He, & Kim, 2015) and the benefits of a low cost yet highly functional system are added to a Honeynet framework. Another APT detection methodology similar to this research is presented in (Jasek et al., 2014), however this research differentiates from others by importing HUMINT Clandestine Network strategy in it. The advantages and ramifications of such systems are discussed in Section 5. Significant work have been done on analyzing the collected data by honeypots in (Ghourabi, Abbes, & Bouhoula, 2014; Prathapani, Santhanam, & Agrawal, 2013; Zhan, Xu, & Xu, 2013). A design for installing honeypots on industrial control systems via proxy servers is discussed in the paper of (Winn, Rice, Dunlap, Lopez, & Mullins, 2015) and on small scale organizations using open source tools is discussed in the paper of (Singh, Sharma, & Singh, 2013).

4. Proposed honeypot network

A Clandestine Cell is a type of an intelligence organization where a cell only knows the immediate superior and the associated members of itself. This kind of organizational structure is used by intelligence agencies throughout the world to provide a security against a breach, thus ensuring the safety of the members. Compartmentalization in clandestine cell networks provides minimization of damage due to exposure of one of the cells or honeypots. The visible part to the attackers of the proposed network is only the cells that are in direct contact with the attackers. Therefore, in case of exposure, removal of one single element of the network is A Clandestine cell communicates in indirect passive methods. They are; Dead-drops (Letter-drops or Mail-drops), Live-drops and Steganography.

- A Live-Drop is a technique where couriers are used in order to deliver messages or items and at the drop site, the receiver waits to secure the package.
- In Dead-Drops, one of the members places a message or item in the drop site and leaves a certain message to another location as an indicator to the receiver of the message. After some time later which is unknown to the dropper, the receiver recovers the package. Known as “the safest form of communication” (Codevilla, 1992) in between the case officer and the agent who is run by him in the intelligence circles during the cold war years, the dead-drops would be used in substitution for knowledgeable human beings wherever feasible (Cooper & Redlinger, 1990).
- Steganography is the art of hiding information in an ordinary file, picture, text or other kinds of media.

In this research, Dead-drops and Steganography is implemented for providing the clandestine communication of honeypots. A communication in a cell is illustrated in Figure 1.

A cell member is a host to a virtual machine having a high interaction honeypot. High interaction honeypots are implemented as Linux servers (Ubuntu 14.04 Server Edition). Vulnerable services are implemented in accordance with the APTs characteristics as shown in Table 1.

A honeypot cell member has two IP addresses shown as IP1.1 and IP1.2 in the Figure 1. These IPs are used when an attack is being conducted. The first IP address is the indicator location where a cell member sets a predefined flag to notify the cell master that it has the log files of an attack. Second IP is the location where the log files are uploaded in a crafted image prepared with steganography. For steganography implementations, Steghide is used in this research (Hetzl, 2003).

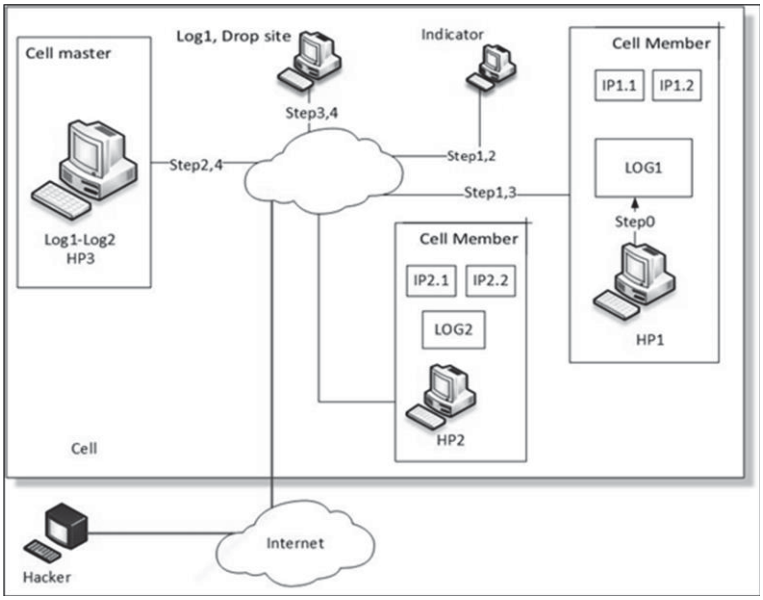


Figure 1: A cell of the proposed network

The cell master is responsible of polling the indicator hosts periodically in a predefined time interval. When it sees a flag is set, then it connects to the second corresponding IP location which is a drop site of the cell member, and extracts the crafted image from the virtual honeypot immediately and sends it to the cell master. The master of the cell, after successfully extracting the information, randomly selects another two different IP addresses from a predefined pool and leaves them to the drop site. The sequence diagram for these two flows are

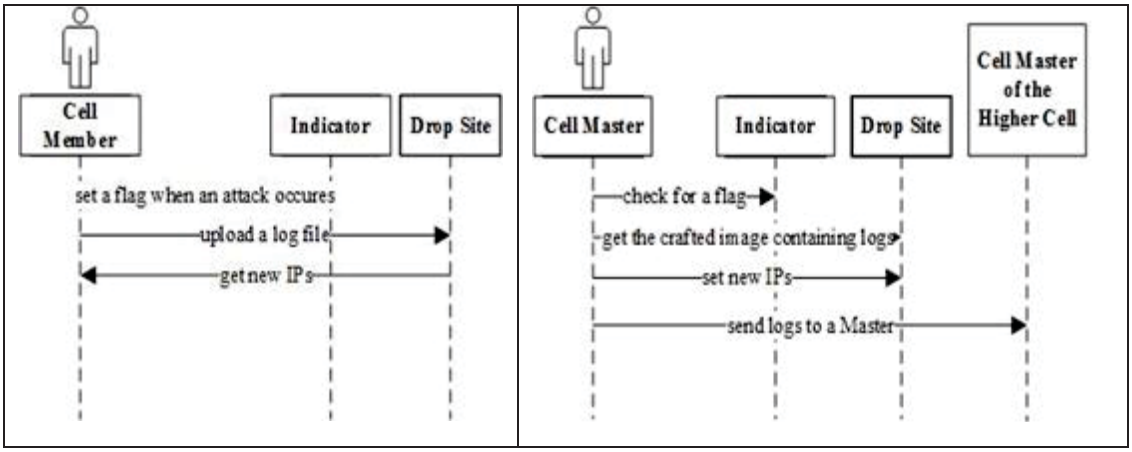


Figure 2: Sequence diagrams for the communication of a cell member with the cell network (a) and cell masters communication with the cell network (b).

Depending on the size of the network, this clandestine cell approach can be increased in levels. Figure 3 illustrates such wide networks compartmentalized by the cells. On top of the system, a Command and Control (C&C) database is installed where all the information about attacks and attackers are saved.

The log files includes source IP addresses of the attacks, ports that are under attack, type of the attack, all the network connections with the honeypot that are established after taken control and to identify the type of the APT, how the communication is achieved. All these logs are compressed and inserted in a JPEG file.

5. Discussions and conclusion

As a consequence of zero-day vulnerabilities and the concept of APTs, it is assumed that the system eventually will be possessed. By compartmentalizing the overall honeypot network design with this approach, these advantages are achieved:

- The log files are extracted from the cell member and not from the honeypot itself. Thus, when the honeypot is possessed, the chances of fooling the attacker is higher as there are no traces of an IDS process or service.

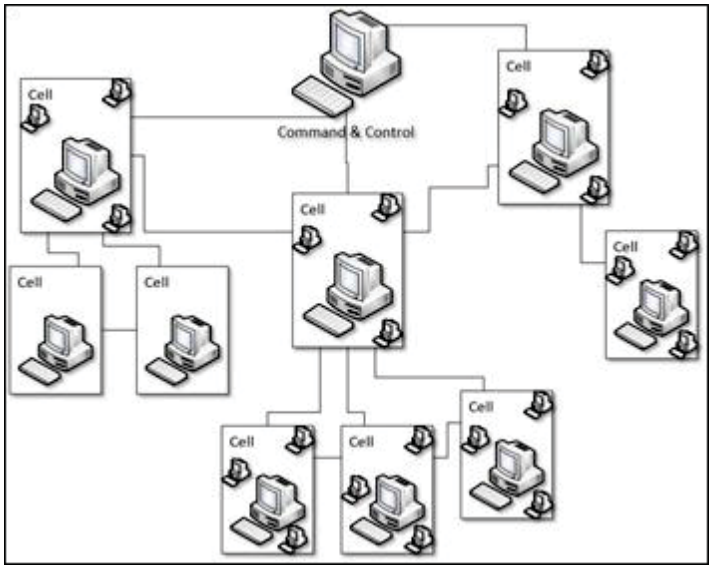


Figure 3: Overall schema for the clandestine network

- By utilizing the indicator and drop site hosts, the communication of the cell member and cell master can never be exposed.
- Even a cell member or a complete cell is possessed, the system will continue to work and the analysis of the attacks can be disseminated by the C&C in real time.
- When compared with a decentralized honeypot network, this approach provides the advantage of automated collection of logs and identifying the diffusion of the attacks.

The main disadvantages of this system are the high cost of installing, maintaining it and overall complexity of analyzing the log files from all cells.

Being the tools of a spy craft, the dead-drops and honeypots were extensively utilized in human intelligence operations during the cold war years, and are still being used contemporarily as well, albeit in a different space defined by computers and other means of communication devices collectively known as the cyberspace. Some of the present day applications of dead drops and honeypots are thus being delineated in this paper.

References

Bencsáth, B., Pék, G., Buttyán, L., & Félegyházi, M. (2012). The Cousins of Stuxnet: Duqu, Flame, and Gauss. *Future Internet*, *4*(4), 971–1003. http://doi.org/10.3390/fi4040971

Chavez, R., Kranich, W., & Casella, A. (2015). *Red October and Its Reincarnation*. Retrieved from https://www.cs.bu.edu/~goldbe/teaching/HW55815/presos/redoct.pdf

Codevilla, A. (1992). *Informing Statecraft: Intelligence for a New Century*. New York: Free Press.

Cohen, F. (1998). A Note on the Role of Deception in Information Protection. Retrieved from http://all.net/journal/deception/deception.html

Common Vulnerabilities and Exposures. (1999). Retrieved February 19, 2016, from https://cve.mitre.org/about/index.html

Cooper, H. H. A., & Redlinger, L. J. (1990). *Catching Spies*. USA: Bantam.

Digby Diehl, & Clarridge, D. R. (1997). *A Spy For All Seasons: My Life in the CIA*. Scribner.

Doherty, S., & Krysiuk, P. (2011). *Trojan.Taidoor*. Retrieved from http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/trojan_taidoor-targeting_think_tanks.pdf

Earley, P. (1997). *Confessions of a Spy: The Real Story of Aldrich Ames*. Blackstone Audiobooks; Unabridged edition.

Eisenberg, T., Gries, D., Hartmanis, J., Holcomb, D., Lynn, M. S., & Santoro, T. (1989). The Cornell commission: on Morris and the worm. *Communications of the ACM*, *32*(6), 706–709. http://doi.org/10.1145/63526.63530

Falliere, N., Murchu, L. O., & Chien, E. (2011). *W32.Stuxnet Dossier*. *Symantec-Security Response* (Vol. Version 1.). Retrieved from https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf

Fanfara, P., Dufala, M., & Radušovský, J. (2013). Autonomous hybrid honeypot as the future of distributed computer systems security. *Acta Polytechnica Hungarica*, *10*(6), 25–42.

Fireeye. (2014). *Assesing Damage and Extracting Intelligence*. Retrieved from https://www.fireeye.com/content/dam/fireeye-www/global/en/current-threats/pdfs/rpt-poison-ivy.pdf

Ghourabi, A., Abbes, T., & Bouhoula, A. (2014). Characterization of attacks collected from the deployment of Web service honeypot. *Security and Communication Networks*, *7*(2), 338–351. http://doi.org/10.1002/sec.737

Hetzl, S. (2003). Steghide. Retrieved February 19, 2016, from http://steghide.sourceforge.net/

Information Warfare Monitor. (2009). *Tracking GhostNet*. Retrieved from https://www.nsi.org/pdf/reports/Cyber Espionage Network.pdf

Jasek, R., Kolarik, M., & Vymola, T. (2014). Extended system of honeypots to detect threats, *8*.

Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security and Privacy*, *9*(3), 49–51. http://doi.org/10.1109/MSP.2011.67

Neumann, J. Von. (1969). Theory of self-reproducing automata. *Information Storage and Retrieval*. http://doi.org/10.1016/0020-0271(69)90026-6

Prathapani, A., Santhanam, L., & Agrawal, D. P. (2013). Detection of blackhole attack in a Wireless Mesh Network using intelligent honeypot agents. *Journal of Supercomputing*, *64*(3), 777–804. http://doi.org/10.1007/s11227-010-0547-3

Sancho, D., Torre, J. dela, Bakuei, M., Villeneuve, N., & McArdle, R. (2012). *IXESHE: An APT Campaign*. Retrieved from http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_ixeshe.pdf

Saud, Z., & Islam, M. H. (2015). Towards Proactive Detection of Advanced Persistent Threat (APT) Attacks Using Honeypots. *Proceedings of the 8th International Conference on Security of Information and Networks*, 154–157. http://doi.org/10.1145/2799979.2800042

Singh, G., Sharma, S., & Singh, P. (2013). Design and Develop a Honeypot for Small Scale Organization, *2*(3), 170–174.

Spitzner, L. (2003). Honeypots: Catching the insider threat. *Proceedings - Annual Computer Security Applications Conference, ACSAC*, 170–179. http://doi.org/10.1109/CSAC.2003.1254322

Symantec. (2011). *W32.Duqu The precursor to the next Stuxnet*. *Symantec Security Response* (Vol. version 1.). Retrieved from https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_duqu_the_precursor_to_the_next_stuxnet.pdf

Varma, R. (2010). *McAfee Labs: Combating Aurora*. Retrieved from https://kc.mcafee.com/resources/sites/MCAFEE/content/live/CORP_KNOWLEDGEBASE/67000/KB67957/en_US/Combating Threats - Operation Aurora.pdf

Virvilis, N., Gritzalis, D., & Apostolopoulos, T. (2013). Trusted computing vs. Advanced persistent threats: Can a defender win this game? *Proceedings - IEEE 10th International Conference on Ubiquitous Intelligence and Computing, UIC 2013 and IEEE 10th International Conference on Autonomic and Trusted Computing, ATC 2013*, 396–403. http://doi.org/10.1109/UIC-ATC.2013.80

Winn, M., Rice, M., Dunlap, S., Lopez, J., & Mullins, B. (2015). Constructing cost-effective and targetable industrial control system honeypots for production networks. *International Journal of Critical Infrastructure Protection*, *10*, 47–58. http://doi.org/10.1016/j.ijcip.2015.04.002

Wright, P., & Greengrass, P. (1988). *Spycatcher CST*. New York, U.S.A.: Dell Pub Co.

Zhan, Z., Xu, M., & Xu, S. (2013). Characterizing honeypot-captured cyber attacks: Statistical framework and case study. *IEEE Transactions on Information Forensics and Security*, *8*(11), 1775–1789. http://doi.org/10.1109/TIFS.2013.2279800

Zhang, W., He, H., & Kim, T. hoon. (2015). Xen-based virtual honeypot system for smart device. *Multimedia Tools and Applications*, *74*(19), 8541–8558. http://doi.org/10.1007/s11042-013-1499-4

PhD Research Papers

The Quincy Wright Model: Postmodern Warfare as a Fifth and Global Phase of Warfare

Sakari Ahvenainen
Finnish National Defence University, Helsinki, Finland
sakari.ahvenainen@kolumbus.fi

Abstract: This article introduces and extends a less known model of history of warfare. It is based on one of the largest academic research projects ever carried out on warfare. It also evaluates the extended version for postmodern warfare as an application and a prediction of the model. The original work is a model of the evolutive phases of warfare. It was published in 1942 as a part of a book "A Study of War", in two-volumes by Quincy Wright, a pacifist and professor of law. The evolutive phases of warfare were animalistic warfare (up to 50,000 BC), primitive warfare (after 50,000 BC), historical warfare (after 3000 BC) and modern warfare (after 1500 AD). New information technology made it possible to create bigger human organizations and caused bigger wars in these phases. These technologies and their levels were protolanguage (clan), language (tribe), writing (state) and printing press (cultures?). This article points out that cybernetics and systems theory support the interpretation of the Wright's model of history of warfare although both cybernetics and systems theory have been introduced after the publication of Wright's book. The presented model predicts the warfare of our time in a surprisingly accurate, though general manner. The next mega phase of warfare, the postmodern warfare, will have at least the following qualities: It will introduce a new kind of communication technology following the printing press. Its main organizational level is global. Its main accounts are not science and technology, as they were the main accounts for the fourth phase (1500–2000 AD), but something else. As a new mega phase, its application and patterns are new and emergent and thus a surprise to us all. It will be a short phase which will only last some 50 years. Its ongoing transition period will include change and chaos and a rearrangement of everything we know.

Keywords: postmodern warfare, history of warfare, Quincy Wright, global warfare, history of information technology, evolution of human culture

1. Introduction

This article presents, evaluates, extends and uses a model of evolution of warfare. This model has some predicting abilities. This article asks, what we can say about the postmodern warfare based on this predicting model.

According to e.g. Wright (1965, pp. 27, 36, 377), McNeill & McNeill (2006, p. 287) Bousquet (2007, p. 9), Fukuyama (2011, pp. 24, 85, 94, 111, 113, 118, 440) and Mirazón Lahr, et al. (2016) warfare has been, is and will be an enormous human enterprise. McNeill & McNeill (2006, p. 275) states that we have lived in this evolutionary human enterprise now for about 500 years in the era of modern science. Bousquet (2007, p. 24) maintains that it has in general given us the power to rule the human environment (nature) and to advance our well-being.

Do we have these kind of scientific theories or models of warfare that advances our understanding of the environment of contemporary or even future wars? Can e.g. Clausewitz or Sunzi say something specific about the future war? No, not at least specific. To contribute our discussion on this important subject as a part of postmodern war, this article presents such a theory or model.

According to e.g. Berlin (1988, pp. 9-10) and Roland (1997) there is a systematic, scholarly and pioneering study to understand warfare in the above sense. It is a book of two volumes and 1637 pages written by Quincy Wright (1965), a pacifist and professor of law. It was first published in 1942 and second edition came in 1965. Wright's book, titled "A Study of War", was a book of summary of 66 scientific studies. The project lasted sixteen years, from 1926 to 1942. The first 408 pages of this book presents a coherent and holistic history of warfare from circa 1,000,000 BC to 1962 AD. Those pages are the core of this article.

The outline of this article is as follows: Section 2 after the Introduction presents a model of four mega phases of history of warfare and the three transitions between them based on Quincy Wright's book. This model is referred from here on as the QW_Model. Section 3 integrated the features of the three transition phases. Section 4 presents the prediction of the fifth mega phase of history of warfare based on the section 3. Section 5 and 6 are short discussions about systems theory and cybernetics and if the QW_Model fits with them. Section 7 is a discussion of the whole article and section 8 is conclusions of the article.

2. The Quincy Wright's model of mega history of warfare and its critics

2.1 The QW_Model

Next we will present the QW_Model, but only generally, concentrating on the levels of human organization, communication technologies and general characteristics of the society and its warfare. Acoording to Wright (1965, p. 30) the four mega phases of history of warfare were animal, primitive, historical and modern warfare. Wright (1965, p. 27, 37) considers these phases as an outgrowth of its predecessor and that new information technologies made these new phases possible. Both of these statements are fundamental for this article. Basic attributes of this development are summarized in table 1. It is important to realize that this model is cumulative. It means that in a new phase, the old is *always* still there.

Table 1: The QW_Model: Mega history of warfare (and societies) according to e.g. Quincy Wright (Wright, 1965, pp. 25-405) (Roland, 1997) (Buzan & Little, 2000) (McNeill & McNeill, 2006) (Merlin, 1991)

	Beginning of the phase	**Organizatio-nal size[3]**	**Type of the society**	**Explanatory Discipline**	**Significantly new**
Protolanguage	Before ca. 50,000 B.C.[5]	Extended family, clan (kin)	Animalistic	Psychology	Fire and primitive stone tools, modelling of the society by mimics, first human expansion out of Africa, …
Language	Ca. 50,000 B.C.	Tribe[4]	Primitive	Sociology	Advanced stone tools, trade and large game hunt, modelling of the outside universe by myths, later agriculture, nomadism, …
Writing	Ca. 3.500 B.C.	States	Historical	International law, politics, economy	Metals, organization, discipline, standing armies, class society, memory of the elite outside the brains, …
Printing	Ca. 1.500 A.D.	Culture[2]?	Modern	Science (Technology)	Modern states and science, universities, democratizing of information, guns and explosives, compass, rule of the seas, clock, later industry, steam, petrol, use of air, space and EM-spectrum, speed of light, …
Global computer technology[1] (GCT)	Ca. 2.000 A.D.	Global[6]?	Postmodern	Cybernetics? Complexity Theory?[7]	Computer and computer networks, the Internet, … (more, see section 4)

[1] = This phase is a prediction from Wright's model. See section 2.2. and 4. - 6. for details.
[2] = Wright (1965, p. 37, 677) called this phase "world community" and considered that now the human race as a single unit was possible (p. 376).
[3] = Highest organizational unit of human society.
[4] = Fukuyama's (2011, p. 53) timing: 9000 BC. Marean's (2015) timing: 160,000 – 70.000 BC.
[5] = Homo erectus, homo heidelbergensis.
[6] = Wright considered the previous phase already as global.
[7] = According to Bousquet (2007), chaoplexic, combination of chaos and complexity.

The QW_Model is a general model which has been stripped out of details, local adaptations, setbacks and fluctuations. One can talk here even about theoretical history based on Bertalanffy (1968, pp. 109-119) and his seminal book of "General Systems Theory".

We will complete below every mega phase with some observations from some newer books. First one is professor Donald Merlin's (1991) book "Origin of the Modern Mind - Three Stages of the Evolution of Culture and Cognition". Second one is professor Barry Buzan's and professor Richard Little's (2000) book "International Systems in World History – Remaking the Study of International Relations". Third one is professor Francis Fukuyama's (2011) book "The Origin of Political Order - From Prehuman Times to the French Revolution".

To start, Fukuyama (2011, p. 24, 30, 62) and Wright (1965, p. 36, 373) state that humans have *always* lived in social groups and had wars or conflicts. According to McNeill & McNeill (2006, p. 22) there has *always* been cooperation as well as competition in human evolution. Fukuyama (2011, p. 34) states that for both of them big brains and human minds were important.

According to Wright (1965 pp. 42-52) *the phase of animal warfare* was before 50,000 BC. It was based on communication techniques preceding language, on clans and extended families. Merlin (1991, pp. 129, 161) maintains that it can be described as the era of protolanguage at its closest phase to the modern language.

According to Merlin (1991, p. 200) the major application of this mimic protolanguage was its ability to model the whole prehuman hominid society. Merlin (1991, p. 198) states also that its cultural applications included toolmaking, fire, coordinated seasonal hunting, rapid adaptation to climate and ecology, intricate social structures and primitive ritual. Furthermore, Merlin (1991, p. 200) concludes that it was a new vehicle for social control and coordination.

According to Wright (1965, pp. 53-100) *the phase of primitive warfare* was after 50,000 BC and Fukuyama (2011, p. 46) agrees. It was based on modern language, tribes and chiefdoms. According to Merlin (1991, pp. 273-4, 308) and Buzan & Little (2000, p. 395) it was limited to the biological memory of man and his brains.

A new research of Marean (2015) puts the start of this phase between 160,000 and 70,000 BC. It also connects it with quite many things, first to *cooperation* with unrelated individuals (hyperprosociality), then *unification* of about 20 clans of about 25 persons to a small tribe of 500 people, a kind of superpower of that time. Other pieces of this development were new advanced projectile weapons, the change of the environment and climate and the first abundant and predictable food source which was worth defending, coastal shellfish beds.

Accordning to Wright (1965, pp. 101-165) *the phase of historical warfare* was after 3000 BC. It was based on invention of writing and states. Merlin (1991, pp. 278, 285, 288- 9) and Buzan & Little (2000, pp. 172, 397) put the earliest writing back to the first city-states and their trade to about 4000 BC. A bigger size of human organization was possible with writing, larger than the primary group, most importantly states.

According to Wright (1965, pp. 166-371) *the phase of modern warfare* began after 1500 AD. It was based on invention of the printing press, on age of discovery and on a global level, at least according to Wright (see section 2.2. Discussion …):

> *"The age of discovery less than five centuries ago marked the beginning of new epoch in human history. … The epoch since the age of discovery, the first of genuine worldhistory, was initiated by the invention of printing in the West …" (Wright, 1965, p. 376)*

2.2 Discussion about the model

In the QW_Model the start is "easy and obvious": protolanguage creates clan, language creates tribe and writing creates state. The next ones, cultural and global are more difficult, also for Wright. Wright (1965, p. 37) speaks about global level (world-community) as the era of the printing press. But global computer technology (GCT) is obviously at least as big a change as the printing press was. What is "left" for GCT, if the printing press already presents the global level? If we separate the ages of the printing press and GCT, as is done in table 1, we get also a good explanation for the rule of the west from 1500 AD to our times. It was the age of technology and science of a culture. The winning culture was the West based e.g. on the printing press. The point is that according to McNeill & McNeill (2006, p. 267), the West was some *hundreds of years* (1450 – 1800) the sole user of the printing press technology. Roland (1997) gives some general critiques of the QW-Model.

3. Integrating: What happened in all these three transition periods?

In the QW_Model there are four phases and three transitions between them: protolanguage, language, writing and printing press. In every transition the following changes have happened (see table 1):

- A new communication technology has been created. It is always bigger and more technological than the previous one.
- It was more open in every new phase. According to Merlin (1991, pp. 160, 289, 296-7, 342, 344, 355, 357) protolanguage, e.g. mimics opened the internal information of human brains to other members of the

community, language opened the past and the future, writing opened information to people in different time and place and printing press from the elite to the masses.

- It has made bigger organizations possible (see section 6). These bigger organizations have bigger technology.
- Bigger organization (society) made more efficient specialization and division of labor possible, as Fukuyama (2011, p. 89) and Derex et al (2013) also suggest.
- Bigger organizations also lead to concentration of people, from small bands of people to village, towns, cities and megacities. This meant more contacts between peoples and their ideas. This is also the basic quality of language, writing and printing press.
- Bigger organizations caused also the rule of new bigger organizations over the smaller ones as Marean (2015) and Fukuyama (2011, p. 81) also suggest.
- Fewer but bigger organizations meant according to Bertalanffy (1968, p. 48) and the systems theory notion "oligopoly" also stronger rivalry, violence and wars.
- More contacts and better and more powerful technology meant shorter and shorter epochs. In the QW_Model it meant about 10-fold decrease in every step starting from about 500,000 years (homo heidelbergensis). Buzan & Little (2000, pp. 405 - 6) has about the same time division.
- According to Wright (1965, pp. 40 - 1) new emergent explanatory discipline of the new era has been born at the new level, but all previous explanatory disciplines did survive as part of the new level at the lower, previous levels, although the older ones were modified and influenced by the new explanatory discipline.
- From concrete towards abstract. This is obvious already in communication technology. A written word is a written abstraction of the spoken abstraction of the spoken word which is an abstraction of its concrete target in the real world. According to Merlin (1991, pp. 269-360) one big milestone in the development of abstraction was the birth of the theoretic mind in Greece about 3000 years ago.
- All this meant an increase of complexity. If there were many of these new units, to unify them one needed again among other things a new kind of communication technology. This brings us back to the starting point.

4. The prediction: The fifth mega phase, postmodern warfare

We know from the previous section what changes happened in all of the previous three transitions. We will present them next with a new prediction (in *italics*) for the fifth phase.

A new communication technology has been created:

Prediction 1 (P1): After printing press the hypothesis is GCT, mostly but not exclusively the Internet. It is an application of the notion "systems of systems".

It was more open in every new phase:

P2: Some parts of the global information are available in the era of GCT to almost every person, everywhere, all time, on the move. Good examples are Google, cloud services, eGovernment, homepages, Facebook and Twitter.

It has made bigger organization possible:

P3: The hypothesis is the global mankind, if there were earlier about eight cultures. New kind of organizations, especially on global level, are part of this development, e.g. Non-Government Organizations (NCOs) like Wikileaks and Anonymous. They increase also the complexity of actions for global actors. Bigger cultural and global organizations mean also less power for the states, as Buzan & Little (2000, pp. 365 - 7) suggest by their notion "the post-Westfphalian era".

Bigger organization (society) made more efficient specialization and division of labor possible:

P4: We live in an era of global trade and China as a factory of the world. According to Levinson (2010, p. 268), *over two thirds of shipping containers crossing the oceans have components or "intermediate goods," inside them, so manufacturing that occurs between (!) factories.*

Bigger organizations meant also concentration of peoples and their ideas:

P5: According to McNeill & McNeill (2006, p. 413) *over half of people lived 2001 in cities. This means more megacities and the Internet is the "megacity" of ideas.*

Bigger organizations meant also the rule of new bigger organizations over the smaller ones:

P6: For information era an interesting example is the UKUSA – agreement between the United States, United Kingdom, Canada, Australia and New Zealand to cooperate on global signal intelligence. And of course NATO and EU.

Fewer organizations meant also stronger rivalry, violence and wars:

P7: The First and the Second World War were examples of this development, also the Cold War and maybe the nuclear war that we have not experienced, the ultimate example.

More contacts and better and more powerful technology leaded to shorter and shorter epochs:

P8: According to the QW_Model, the previous phase lasted about 500 years (1500 – 2000), so the next postmodern phase will last some 50 years. A candidate for its star is the birth of the Internet.

New, emergent explanatory discipline has been born:

P9: The explanatory discipline of the previous phase were science and technology (of the cultures). They will survive, but what could be the next ones? A choice is information and its science, cybernetics. Information technology is a part of computers and of media and a very big and growing industry, not to forget Information and Cyber Warfare. It is interesting that according to Floridi (2011) and (2014), information philosophy is also on the rise. Bousquet (2007) suggest that could the explanatory discipline be chaoplexic, combination of complexity and chaos theory? Bigger means in principle more complex. We point out here also that according to Pagels (1989) computer is a research tool to study and to understand complexity. This hypothesis combines theories of information, computer, complexity and chaos.

From concrete towards abstract:

P10: According to Wiener (1948, p. 132), information is an abstract thing, not matter, nor energy. Still it connects all (cybernetic) systems together (see section 6). So if the postmodern era is era of information, it is also the era of abstraction. Arquilla & Ronfeldt (1999) suggest that in strategy this means e.g. soft and open noopolitik, globe-spanning realm of the mind.

All this meant increase of complexity:

P11: After global mankind there is no human expansion potential on Earth. We have to look for the other planets, even stars and new communication technology between the stars.

To summarize, these predictions describe some aspects of our postmodern era. An illustrative summary of key words of the postmodern based on this section is presented in section "Conclusions".

5. Does the QW_Model follow systems theory?

We will first list in brief in this section what are the basic notions of open systems and systems theory and then if the QW_Model follows these basic notions, which are according to Bertalanffy (1968, pp. 27, 49, 55, 66, 96, 121, 215, 219):

- open system and its environment and the borders between them and their relationships (inputs and outputs)
- first the whole of the system and then its elements (parts) and thirdly their relationships, including competition between the elements of the system

- emergence
- system levels, hierarchy

Only emergence and system borders are presented here below in more details, as they have the most obvious connections to the QW_Model. Emergence is the key to understand the essence of systems and our world. According to Skyttner (2007, p. 65) and Pagels (1989, p. 223) emergence creates systems levels and even new theories to explain them. Emergence is an integrated effect of the parts of the system. Bertalanffy (1968, p. 55) maintains that It has a stable form, new kind of pattern which none of the parts have. In the QW_Model emergence is most distinguishable in the explanatory disciplines (table 1), which are and should be new to every bigger level of human organization and its war. This is one reason to separate cultural and global level and technology-science and information-complexity in table 1.

Borders work also quite well in the QW_Model. Human organizations were first small with simple technologies and had a "influence border" between them. Influence borders were a communication problem but also problems of limited means in economy, military and transportation and of geography, e.g. mountains, deserts, rain forests and oceans.

6. Does the QW_Model follow cybernetics?

We will first list in brief in this section what are the basic notions of cybernetics and then if the QW_Model follow these principles of cybernetics, which are according to Bertalanffy (1968, pp. 43, 150), Turchin (1977, p. 17, 25-6), Skyttner (2007, p. 91), Wiener (1948, pp. 160 – 1) and Gleick (2011, pp. 355-372):

- sensor, decision making unit, effector, setup value, borders and feedback loops are the basic parts of cybernetics systems
- information as an abstract difference but always connected to a physical (cybernetic) system
- information as the part which connects organisms (cybernetic systems) to a holistic entity, so means of control
- cybernetic system as a system which can process information by decision making unit (e.g. nerve net) from sensor to effectors to influence its environment, to convert a sensor situation into an effector action
- new e.g. bigger systems need new kind of information systems, because every (communication) systems has its operational limits.

Only the following cases are discussed in more details below, as they have the most obvious connections to the QW_Model: communication as glue and means of control and new level and its new communication system.

Wiener (1948, p. 156) considers that it is specifically the communication between the parts of the organization which makes it intelligent, even if it consists of simple parts, for example honeybees and ants.

When the size of the organization grows beyond the possibilities of its information technology, new means for acquisition, utilization, storage and dissemination of information are needed. The main point between cybernetics and the QW_Model is that language, writing, printing press and GCT make the growth of the human organization possible, from clan, tribe, states, cultures to global mankind. That is also one reason why historical warfare (3000 BC to 1500 AD) did not develop to a new level after tactics of primitive warfare of the tribes. There were no new communication means to control the bigger areas of the greater battles (operations) of a state.

To end this section, we will point out that according to Skyttner (2007, pp. 416-7) the famous John Boyd's OODA-loop (Observation - Orientation – Decision – Action) is a cybernetic system very much in line with the features outlined in this section.

7. Discussion

We have extended the QW_Model to the next phase of warfare and evaluated it for postmodern. The basic problems are here first the complexity and grandness of the whole. Secondly, there is by definition few evidences from the prehistory. Cultural and global phases are also a bit difficult in this model.

This article is based mainly on one source, although special, even seminal. From the new research e.g. Merlin (1991), Buzan & Little (2000), McNeill & McNeill (2006) and Fukuyama, 2011) support it. The QW_Model also

follows systems theory and cybernetics at least in some very important features. This conforms to some aspects of both systems theory and cybernetics *and* Wright's work.

It is found in the article that the three transitions which have happened in the QW_Model have many of the same changes. This gives a solid base for prediction for the next phase. It is also found that these predictions have an interesting amount of common features with our time and its warfare. The model may give even some strategic thoughts to act in the contemporary world.

The change suggested in this article can be difficult to see at first, because the old, e.g. states are still present as the old has always been (see table 1). Secondly it is difficult to see, because the new is always an emergent form. And lastly, this development is not straightforward as recent developments in Russia, Ukraine and EU show. But has it ever been in the evolution of these sizes?

8. Conclusions

If we and Wright are right here, it is imperative to us to understand, how much we owe as human beings to each other. It is also important for us that the current phase of this process will work to keep or even to advance our well-being. Humans would not survive long alone and would survive in a closed extended family unit, but could not support technology higher than Stone Age, as Derex, et al. (2013) suggests. McNeill & McNeill (2006, p. 316) suggests that also the late Tasmanians were a sad example of this.

If the reader finds the essence of this article usable and interesting then we want to contribute that feeling to the scale of our and Wright's method: If one looks far, one sees more, not the details. Buzan & Little (2000, p. 385) point out that this was also the main starting point of their new theory.

The QW_Model is based on changes in information technology. This is of course interesting for our era of information and era of information based warfare. It highlights likewise the importance of cybernetics as an information based theory, even in warfare, as the OODA-loop suggests.

The QW_Model gives also a different view to global computer technology (GCT). In the language of the QW_Model it is a tool to rule the complexity of global mankind and to enable a new level of human organization.

If we accept the QW_Model, we accept that history repeat (!) itself, not in detail, but in system principles.

The main conclusion about postmodern warfare is that the chaos and complicity of our world (2016) has a good general explanation in the QW_Model. We live in a transition period between two mega phases, culture and global mankind, which may even overlap.

This model says also what is *not the most important* in postmodern warfare; *science and technology*, as they were the old "explanatory disciplines" of the previous cultural phase. Now it is something else and we have only some hypotheses at the moment. They are cybernetics and complexity theory *on a global scale*. So the new military strategy, strategic communication or noopolitik are close calls.

To start the ending, we present some key words of the postmodern society and its warfare *according to this article*:

- global level in general
- its numerous actors, e.g. states, NCOs, (international) firms, crime, terrorism and infectious deceases
- global trade and division of labor
- internet, e.g. cyber security and internet of things
- other global communication systems, e.g. ocean crossing cables, satellites, mobile and smart phones and media
- computers and computer networks everywhere, also as as new information actors besides human
- information in many forms, e.g. Big Data, eMoney, open data and open software
- complexity and rule of complexity

- urban life and war (megacities)
- more well-being
- increased global control
- increased speed
- increased abstraction and of course, first time
- possibility of *global* catastrophes caused by humans, also, sorry to say, in many ways.

To end the ending, one last concluding remark: If we and Wright are right, do not meditate too long, as this change is fast, global and comprehensive.

References

Arquilla, J. & Ronfeldt, D. (1999) The Emergence of Noopolitik – Toward an American Information Strategy, Santa Monica, RAND.

Berlin, R. H. (1988) Historical Bibliography No. 8 - Military Classics, Fort Leavenworth, U.S. Army Command and General Staff College.

Bertalanffy (von), L. (1968) General Systems Theory – Foundations, Development, Applications, 2003 Edition, New York, George Braziller.

Bousquet, A. J. A. (2007) The Scientific Way of Warfare: Order and Chaos on the Battlefields of Modernity, London, London School of Economics and Political Science.

Buzan, B. & Little, R. (2000) International Systems in World History – Remaking the Study of International Relations, Oxford, Oxford University Press.

Derex, M., Beugin, M.-P., Godelle, B. & Raymond, M. (2013) Experimental evidence for the influence of group size on cultural complexity, Nature, 21 November, Issue 503, p. 389-91.

Floridi, L. (2011) The Philosophy of Information, Oxford, Oxford University Press.

Floridi, L. (2014) The Fourth Revolution: How the Infosphere is Reshaping Human Reality, Oxford, Oxford University Press.

Fukuyama, F. (2011) The Origin of Political Order - From Prehuman Times to the French Revolution, Kindle 2011 Edition, New York, Ferrar Straus and Giroux.

Gleick, J. (2011) The Information - A History - A Theory - A Flood, New York, Pantheon Books.

Levinson, M. (2010) The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger, Princeton, Princeton University Press.

Marean, C. W. (2015) The Most Evasive Species of All. Scientific American, August, pp 32-39.

McNeill, W. H. & McNeill, R. J. (2006) Verkottunut ihmiskunta – Yleiskatsaus maailmanhistoriaan, Tampere, Vastapaino. (Original in English: The Human Web. A Bird's Eye View to World History)

Merlin, D. (1991) Origin of the Modern Mind - Three Stages of the Evolution of Culture and Cognition, Cambridge (Massachusetts), Harvard University Press.

Mirazón Lahr, M. et al (2016) Inter-group violence among early Holocene hunter-gatherers of West Turkana, Kenya. Nature, 21 January, pp 394-7.

Pagels, H. R. (1989) Dream of Reason – The Computer and the Rise of Sciences of Complexity, New York, Bantam Books.

Roland, A. (1997) "Technology and War", [Online], American Diplomacy, http://www.unc.edu/depts/diplomat/AD_Issues/amdipl_4/roland.html

Skyttner, L. (2007) General systems theory – Problems, perspectives. 2nd Edition, New Jersey, World Scientific.

Turchin, V. F. (1977) The Phenomenon of Science - a cybernetic approach to human evolution, Abode Reader pdf Edition, New York, Principia Cybernetica Project.

Wiener, N. (1948) Cybernetics: or Control and Communication in the Animal and the Machine, 10th Edition (2000), Cambridge (USA, The MIT Press.

Wright, Q. (1965) A Study of War, 2nd Edition, Chicago, University of Chicago Press.

Decision-Support by Aggregation and Flexible Visualization of Risk Situations

Alexander Beck[1] and Stefan Rass[2]
[1]VW Financial Services, Germany
[2]Universität Klagenfurt, Institute of Applied Informatics, System Security Group, Austria
alexander.beck@vwfs.com
stefan.rass@aau.at

Abstract: The increasing complexity of infrastructures has nowadays resulted in equally complex attack schemes and threat scenarios. While vulnerability detection enjoys sophisticated tool support, the results obtained from the analysis of an IT infrastructure are often difficult to work with unless they are properly "aggregated". Especially for decision makers, a precise and concise view on the risk situation is inevitable for effective risk management. This work is therefore dedicated to automated risk aggregation and visualization, to provide a decision maker with an easy to interpret overview that highlights the most important spots where action is demanded. As our showcase example, we will use the common vulnerability scoring scheme that measures security in terms of 14 different scores on a nominal (qualitative) scale, which are based on expertise and experience. Giving the set of scores individually for a large number of components quickly renders the resulting report almost infeasible to read or understand. Consequently, it is often the security manager's duty to compile the data into a more accessible form, and to provide a high-level overview. However, most practical such condensations block the reverse way of "zooming into" the risk picture, and thus hide partial aspects of potential importance from the big picture. It is therefore demanding to keep such reports in a form of controllable granularity, and which allows to get more than a brief summary. We achieve this by doing hierarchical risk aggregation using a neural network (to somewhat mimic human reasoning in this context).

Keywords: threat-analysis, security-assessment, neuronal networks, impact-analysis, visualization, decision support

1. Introduction

Risk control is a core duty of enterprise management. For security, various industrial standards have been compiled from best practices and scientific results, which can be used as a guidance and tool for risk managers. The most prominent such standard is the ISO 27k-Family, which prescribes an overall process, yet leaves the particular details of reach step widely open and up to a selection from best practices. Without digging into too much detail about the risk management process as such, we will hereafter focus on the particular task of risk assessment for an entire infrastructure, which basically is the assignment of a quantitative or qualitative risk measure to a given system.

In detail, the usual process is to start with a risk assessment of particular components and their interconnection to other parts of the system. With known vulnerabilities for each component at hand, the common vulnerability scoring system (CVSS) (see (Mell, Scarfone und Romanosky 2015) and (Joh helps to systematically assign scores to paint a detailed picture about the risk exposure of a component. These scorings are often automatically delivered by scanning tools like OpenVAS or Nessus, based on public data bases like the Common Vulnerabilities and Exposures (CVE) database. This "picture" – at its finest granularity – consists of 14 different scores, whose totality makes up the risk assessment for a specific component. For convenience of the reader, we will explain these scores briefly in section 2.1, thus omitting details now for the sake of stepping forward in the risk management process. Now, given the set of 14 scores per component, and a number of N such components, the risk manager needs to aggregate the total of $14N$ scores into a "useful" risk estimate for the overall system, otherwise, this expectedly long list of scores will hardly be helpful in decision support.

To mitigate this issue and to boil down a risk assessment reports to a digestible lot, risks are often aggregated by an expert, taking the system description and individual component's assessments to produce an overall assessment consisting of relatively few scores. This is exactly the point where expertise and personal experience is most valuable, and at the same time, there is hardly any automated tool-support or best practice catalogue available that would help an inexperienced user to fulfil this duty. The need for software support has been widely recognized and topological vulnerability scanners can be combined with sophisticated post-processing tools like "Cauldron" (see (Jajodia et al. 2011) and (O'Hare et al, 2008)) for condensing and visualizing results into a more accessible form for decision making. However, these tools usually do not handle matters of risk aggregation, and primarily help with the information mining associated with risk assessments. Matters of graphical representation

and decision support in general, such as we also discuss in section 4, have been subject to a vast amount of prior research and standardization efforts (see (Noel et al, 2015), (Xinlan et al., 2010), (NIST, 2011), or (Kawasaki and Hiromatsu, 2014) to name only a few).

Our main contribution in this work is using techniques from artificial intelligence to "mimic" what an experienced risk manager would produce from a given set of component scores. In brief, we will train a neural network with data collected from manual (human) risk aggregation, so as to get an automated system that does a risk assessment very much like a human expert would do it. This offers a variety of interesting possibilities, since:

- The risk aggregation can be delegated to an even inexperienced person, since the expertise is "encoded" within the aggregation function
- The process becomes applicable to large-scale infrastructures, where manual aggregation would be time-consuming (and thus inefficient)
- Respective aggregation functions can be compiled into software tools that can be given away as general-purpose decision support tools. We leave this direction for future research, and will concentrate on the aggregation function in this article.

Paper Outline: To ease the start on CVSS for the non-experienced reader, we briefly review the core concepts of this risk assessment method in section 2.1, with a subsequent section used to illustrate the risk aggregation problem. The remainder of section 2 is dedicated to a discussion of alternative techniques to justify our chosen approach within the acknowledged related work. Section 3 gets into the technicalities of the risk aggregation issues, which is divided into matters of reducing ambiguities (section 3.1.1) and automating the risk aggregation process itself (section 3.1.2). Section 4 frames the overall contributions in the more general framework of decision making by describing how the risk situation can be dynamically adapted to different needs for a decision maker. Conclusions and future directions of work are outlined in section 5.

2. Related work and contribution

The following sections serve a double purpose, in the sense of describing the context of our research and using this description to point out practical difficulties for which we contribute solutions in this work. Competing solutions to automated risk assessment are discussed in section 2.2 and later.

2.1 A quick look on the common vulnerability scoring system

CVSS is an industrial standard for security assessments and severities. It was developed from the Forum of Incident Response and Security Teams (FIRST) and commissioned by US Homeland Security. We used CVSS version 2 from 2007 in the following.

A general CVSS assessment consists of three metrics that capture different aspects of a vulnerability. The *base metric* contains all information about the vulnerability, including the exploit complexity (efforts to use the vulnerability to mount an attack) and the impacts (effectiveness of an attack mounted using the vulnerability). This already allows to differentiate simple from difficult attacks on the confidentiality, integrity or availability. The *temporal metric* is based on information about the discovery, effective protection measures and the prominence of the vulnerability, e.g., if the vulnerability is long known or quite new, and similar. The *environmental metric* is based on the possible direct and indirect impact that an exploit would have on the surrounding systems/environment. These three metrics together go into the overall CVSS score, where each individual metric is computed from quantitative scores (functions f in Figure 1). The temporal and environmental metrics are optional in the sense that a CVSS score requires at least the base metric, and can only be refined upon the other two.

Once computed, a CVSS vector is a string containing textual descriptions of the scoring under each metric, and may look as follows:

AV:N/AC:M/AU:S/C:C/I:P/A:N//E:F/RL:W/RC:C//CDP:H/TD:H/CR:H/IR:H/AR:M

Therein, the three metrics are separated by a double slash (//), and each individual metric's inputs are separated by slashes, separating "factor : assessment" pairs. For example, the temporal metric (middle block) starts with "E:F", which reads as "Exploitability: Functional", and tells that a working exploit code is available that

successfully mounts attack using this vulnerability. Likewise, "AC:M" in the base metric would read as "access complexity: medium", telling that an exploit is only possible under certain circumstances and takes some effort.

This vector can be convert in a CVSS score, so that we have an easier presentation of the vulnerability. For example a smaller score on a scale from 1 to 10 is less critical then a higher score. The numerical score is computed using formulas specified in the CVSS standard (Mell, Scarfone und Romanosky 2015), and shall not further bother us here.

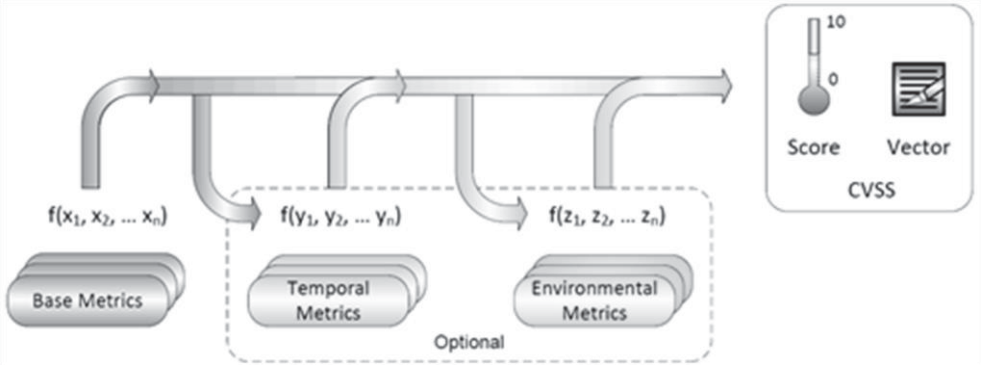


Figure 1: Schematic CVSS computation according to (Mell, Scarfone und Romanosky 2015)

2.1.1 CVSS risk aggregation

Aggregation of risks based on CVSS can thus be done using the numerical score (only) and/or the textual descriptions. Manual aggregation is certainly easier on the textual representations, which avoid loss of information through numerical aggregation. On the other hand, too much such information will at some point make an aggregation difficult again, since the assessments of different components are not always related in an obvious fashion, and dealing with inconsistencies between assessments may create additional difficulties. To practically verify this claim, we examined the quality of replies of risk assessment with several pre-defined vulnerabilities in a hypothetical IT infrastructure. In this study, we observed an inadequate database of information for security experts and their assessments.

Reducing inconsistencies in expert assessments, and their implied difficulties in terms of conflict resolution, is therefore a first step that must be accomplished a-priori to a risk assessment (regardless of whether this is CVSS or another method). One contribution of our work is a concrete method to do this, which is described in section 3.1.1. Roughly speaking, we developed a *meta metric* as an extension for the CVSS metrics. This metrics is based on interviews, studies and other empirical investigations, and shall help experts to get a more detailed picture before formulating their opinion about risk.

2.1.2 Practical issues of CVSS

Somewhat surprisingly, very few practitioners seem to be familiar with CVSS and are able to apply it to practical infrastructures. Along a sequence of interviews conducted with various experts at different companies, we discovered that only a small group has solid practical experience in CVSS. These findings were based on a questionnaire on six scenarios with two vulnerabilities each, where each expert was asked to make an assessment about the risk that both vulnerabilities together imply, i.e., do an *aggregate* assessment. The overall return rate was 46 results obtained from 10 experts. In total we have 25% good and representative answers in our first scenario. Roughly speaking, we could identify three categories of experts, relative to CVSS experience:

- Those without knowledge in CVSS
- Those with theoretical knowledge but no practical experiences
- Those with practical experiences

The last group was the smallest in our result, which indicates that CVSS is quite well known yet hardly applied in practice.

Often, we see unstructured little comprehensible decision-making processes in a security department. Much more experience is usually available in bigger security departments, but this does not rule out ambiguous

opinions (like three experts giving five different assessments). For example, an expert with a good network education will primarily look at the network infrastructures, such as open ports and the security protocols. A second expert being a process security auditor, will focus on security processes and regulatory aspects. Another expert, say a forensic guy, will in turn concentrate on software, database changes, protocols and the network connections. So, different expertise yields to different assessments (which indeed is all justified, yet on different grounds). This intrinsic uncertainty and ambiguity in the data that often underlies security risk assessments at the same time rules out many of the purely mathematical techniques of decision making, such as based on game-theory (e.g., (Miura-Ko et al., 2008), (Alpcan & Başar, 2010)); recent advances to deal with such ambiguity on a mathematically solid decision-theoretic fundament have been made, but are so far not practically available (see (Rass, 2015)).

This simple example scenario is already a reason why we need a standardization in our security assessment in practice. Our proposal here is thus to take an existing standardized method (like CVSS), extend it by information items that experts usually take into account for risk aggregation, and use this additional data to create a wider automation towards risk aggregation in the (so still standardized) method.

2.2 Decision trees and fuzzy aggregation

Decision trees are a systematic and often quite intuitive approach to classify objects, or in our case, to systematically reach a risk assessment (see (Xinlan et al., 2010) for one such method). Especially CVSS assessments would provide a set of attributes that can straightforwardly be used with a decision tree. While this may be a useful way to compute a single component's risk, aggregating two or more risks is not trivially possible using that technique. This is an unfortunate consequence of the intrinsic uncertainty in risk assessments, which arise from unknown threats, unknown external influence factors, and similar. These practically hinder the specification of "exact" rules that could give an aggregate risk from individual risks. For the same reason, fuzzy logic applies only with severe difficulties, since experts may be unable to systematically convert their intuitive method of risk aggregation into a set of if-then rules and membership functions.

Our work nevertheless is related to fuzzy aggregation (see, e.g., (Liu et al., 2005)) in the sense of using a neural network instead of fuzzy reasoning, since neural networks – unlike fuzzy logic – can be "trained" in the sense of a systematic fit to data at hand. Comparable techniques for fuzzy logic or crisp decision making (for decision trees) appear either widely heuristic or unavailable so far.

2.3 Other risk aggregation techniques

The simplest yet most common way of aggregating risks is based on the well-known metaphor that a chain is only as strong as its weakest link. In that spirit, we let the highest individual risk determine the overall risk in a given system. Formally, we consider a system as a collection (set) of components with individual risk assessments and take the maximum of all risks as the system risk (maximum principle; cf. (Bundesamt für Sicherheit in der Informationstechnik 2013)). This method indeed enjoys a solid statistical fundament, as it can be taken as a specific application of the upper Fréchet-Hoeffding inequality from Copula theory. Roughly speaking, this relates to our setting as follows (Rass und Kurowski 2013): let us model the components $x_1, x_2, \dots, x_n$ in a system y with Bernoulli random variables $X_1, X_2, \dots, X_n \in \{0,1\}$ that indicate failure of components (say, if the i-th component is down, we set $X_i = 0$), and introduce another random variable Y as an indicator of the overall system y to fail. The probability distribution of Y is determined by the probability distributions F_{X_i} for the variables X_i, and their joint distribution (modelling the failure of Y) can be computed from the marginal distributions by a so-called copula function $F_Y = C\left(F_{X_1}, F_{X_2}, \dots, F_{X_n}\right)$. Sklar's theorem, cf. (Nelsen 1999), guarantees the existence of some function C that satisfies the last equality, and the Fréchet-Hoeffding bound tells that every such function C must satisfy $C(r_1, \dots, r_n) \leq \min(r_1, \dots, r_n)$. Observe that the right hand side is nothing else than the maximum principle, if we interpret the distribution of Y as the likelihood for a correctly functioning system. Thus, the correct functionality is given if and only if all individual components function correctly. This likelihood, as computed by the function C is, however, no larger than the likelihood for any component to fail, which is the maximum principle, only expressed here in statistical terms. Finding the copula function C is possible based on empirical data, which is mostly likely unavailable (as a sufficient amount of security incidents for probably have devastating consequences for a company before a robust statistical estimate could be created.

Nevertheless, the maximum principle is convenient in lack of detailed knowledge about inner system dynamics (which would be described by the function C). Still, it can be quite wasteful of information, as it ignores all available opinions except the most pessimistic one.

In a different view, the copula function C could also be interpreted as an aggregation function, as it takes individual assessments (marginal distributions) as inputs and outputs an overall assessment (the joint distribution). Given the difficulty of using a copula, our contribution in the following is sort of a replacement for the copula function that fulfils the same purpose, although it is conceptually different.

3. Automated risk aggregation

3.1.1 Reducing answer variability – introducing a meta-metric

Ambiguity of opinions is a common issue in security risk assessments, which is partially induced by intrinsic uncertainty about the current threat situation (we cannot prove the absence of all vulnerabilities), but also due to subjective preferences, personal experience, and many other psychological or sociological factors. These influences imply that answers of different experts (depending on their personality, preferences, risk aversion, experience, etc.) are often diverging and occasionally also mutually contradictive.

To mitigate this problem, a standard aggregation technique simply takes the most pessimistic assessment among all (cf. section 2.3), but a better approach would certainly be to a-priori reduce the variability among the answers (as we briefly discussed in section 2.1.2 already).

We can achieve this by "enforcing" people to give more thought out/self-conscious answers to the questions by making them answer some more questions before moving to the CVSS scoring assessment. In other words, by putting people through a prior questionnaire, the expert is a priori forced to think about the infrastructure more carefully, and will later on provide a "more informed" answer when it comes to the CVSS scoring assignment.

This effect has been tested empirically, with the data displaying some notable improvement of the answer's quality (in the sense of ranging in a narrower interval when the a priori questionnaire was done, compared to a pure CVSS survey).

The study was designed as follows: within a hypothetical IT infrastructure, six scenarios were defined, with two vulnerabilities per scenario. As an example, one scenario involved two servers in an infrastructure and an internet entry point to one server. The first vulnerability was in the entry point. The second weakness is in the service that is used by a second server. The problem for the expert is to assess the common risk regarding confidentiality of the information on the second server, in light of both vulnerabilities.

For all scenarios, the given description was a graph model, with nodes corresponding to infrastructure components, and edges modelling the interconnections between them. The available experts were divided in two groups, where the first group was asked to fill in a prior questionnaire before entering the CVSS assessment stage, and the second group was taken directly to the CVSS assessment. The questionnaire for the first group – our meta metric – asked for the following in a multiple-choice fashion:

1. What is the relevant protection target? This could be confidentiality, availability, etc. This is important to fix a priori, since assessments can be quite different related to what the security goal is. CVSS does not ask for this explicitly. Possible choices:

- *Confidentiality*
- *Integrity*
- *Availabiliy*
- *Not defined*

2. Are there redundancies of the nodes under consideration in the infrastructure? That is, are there potentially multiple ways to sneak into the system, some of which might be easier to follow than others? Possible choices:

- *Yes*

- *No*
- *Not Defined*

3. What is the application type of the nodes? This specifies what the application does with the data, like storage, processing, transmission, etc. Possible choices:

- *Input / Output (e.g., node only transmits)*
- *Input / Output / Processing (e.g., node processes input data and forwards it)*
- *Input / Output / Processing / Storage (e.g., node processes input data, store it temporarily and forward it later)*
- *Usage*

4. What kind of data or information is in the nodes?

- *Payload*
- *Meta Information (e.g., network status information)*
- *Not Defined*

5. How are the nodes relevant for the vulnerability? That is, how much of the system can be affected by a problem within a node (can an infection have only local impact, only be relevant for a single application, or may infect larger parts of the surrounding system?)

- *Single (only relevant for one application)*
- *Multiple (possibly relevant for many applications)*
- *Not Defined*

6. Is the node a security relevant node? For example, a firewall, a virus scanner, etc.?

- *Yes*
- *No*
- *Not Defined*

7. Is the node based on proprietary software or firmware?

- *Yes*
- *No*
- *Not Defined*

8. It the node based on open source software?

- *Yes*
- *No*
- *Not Defined*

All these inputs are taken into account by human experts, but are not necessarily queried for a CVSS assessment. Assuming that differences and divergences among expert opinions arise from different people taking into account different information, the meta metric mainly serves to "equalize" the thoughts of experts before going into the assessment. That is, if an a-priori questionnaire enforces people to consider a common set of relevant influence factors, subsequent CVSS assessments are more likely to be consistent (even across different expertise, education and experience).

The benefit of this became visible in our experiment on the comparison between CVSS with vs. without meta-metric. The results are given in terms of range, mean and variance, as summarized in Table 1. Though the sample was insufficient for a statistical confirmation of the lower variance, the data quality became notably better by our extended approach, as the comparison between the rows with and without the meta-metric indicates. Particularly interesting is the observation that outliers in the environmental metric were considerably reduced. Even though the range of the values remained equal for the temporal score, for instance, its variability measured by the variance was substantially reduced. This indicates that the meta-metric seems to be helpful in making an informed score in this third part of CVSS. Looking at the means in each metric, we see that the overall assessment

is only slightly (yet not significantly) altered when the meta-metric is taken into account. Thus, the overall assessment results do not become biased.

Table 1 Comparison of CVSS Scores with and without Meta Metric

	Range (Delta)			Mean			Variance		
CVSS Score	Base	Temp	Env	Base	Temp	Env	Base	Temp	Env
Without meta metric	6.4 – 10 (3.4)	7.8 – 10 (2.2)	5.5 – 10 (4.5)	8.54	7.99	7.69	1.15	1.02	6.26
With meta metric	7.1 – 9 (1.9)	6.4 – 8.6 (2.2)	7.8 – 9.3 (1.5)	8.34	7.87	8.73	0.47	0.46	0.38

3.1.2 Aggregation of risks

In brief, risks (CVSS vectors) are aggregated using a neural network. We chose this technique for its flexibility and ease of "fitting" its output to many given training sets. Those training sets were obtained from empirical user studies, where we divide the whole set of 45 records at a ratio of 75% training data and 25% verification data. The aggregation is always done on inputs of two 14-dimensional CVSS vectors, augmented by the data from the meta-metric, each of which is a set of 8 inputs, obtained from the 8 questions in the meta-metric as described in the previous section. Thus, the overall network has $2 \times (14 + 8) = 44$ inputs, and outputs another 14 CVSS scores. The aggregation then proceeds by invoking the aggregation network on input of further CVSS assessments, until an entire subsystem has been aggregated "bottom-up". Of course, this assumes that a subsystem can be described by an *acyclic* graph, such as depicted in Figure 2. The essential implication of acyclicity is its implied topological ordering, which lets us consider the infrastructure as a tree-model, with a root (node B in Figure 2), and child nodes being ancestors of their parents (up to the root). On a logical level, acyclicity can be interpreted as the absence of cyclic dependencies. Such cyclic dependencies may indeed arise from feedback loops between applications, however, such bidirectional dependencies can be resolved by decomposing this into two models with unidirectional dependencies, for which a risk assessment can be done separately and aggregated afterwards.

Note that the NN aggregation at the upper levels proceeds with respective networks that do not use a meta-metric any more (thus reducing the number of input nodes accordingly). This is justified since the meta metric was already aggregated at the bottom layers, and thus "implicitly propagates" up to the top.

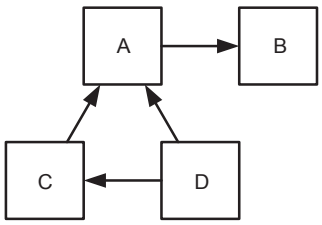


Figure 2: (Logical) Interdependencies between applications for aggregation

The network itself is a perceptron with one hidden layer, a bias node and the hyperbolic tangent as the transfer function. The topology is sketched in Figure 3.

4. Decision Support

4.1.1 Refining and coarsening the risk picture

Let us now return to the model on which the risk aggregation was performed (such as shown in Figure 3). A main advantage of the automated risk aggregation is the possibility to collect information along the bottom-up aggregation to display the aggregated risk at different levels of abstraction (in manual risk assessments, this information is likely to be lost or abandoned). Assuming the system model graph to be acyclic, we can display the network in a tree-like structure, with the "root" being the smallest node in the topological order of the graph (the existence of this order is assured by the acyclicity and easy to compute). Define the distance of a component X as the length of the path from the root to the node X, we can define the n-th layer in the network topology model as a set of components at distance n to the root node. It is then a simple matter to zoom-in or zoom-out of the picture by restricting the picture to any specified number of layers, starting from layer 0, which is the root node. Figure 4 displays an example of three components A, B and C, which are bottom-up aggregated. Letting

AB denote the aggregated CVSS valuation of A and B, and likewise writing ABC for the overall aggregate risk, we end up with three possible views that a decision maker can ask for. Normally, the top view (only layer 0) will be preferred, taking a zoom-in only upon the need to investigate where an inacceptable risk situation originates from (in lower layers).

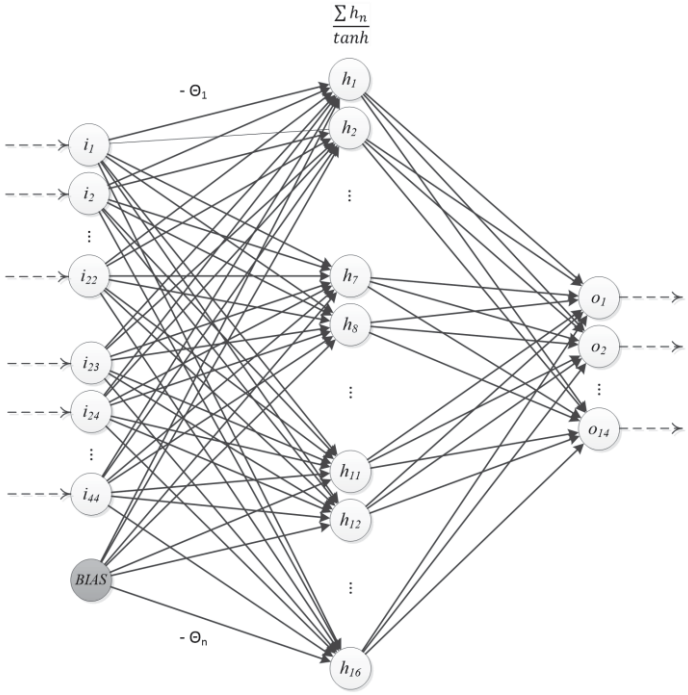


Figure 1 Aggregation network topology (Heidorn 2014)

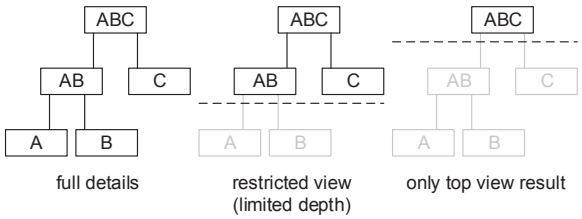


Figure 4: Zooming the risk picture

Similarly, it is easy to augment the system model by geographic information to geo-locate certain problems that become visible in the top level view, say if the overall risk is unacceptably high. The necessary geographic data is herein assumed as available in the system (which appears reasonable, since the devices are part of a fixed installation, and mobile parts of the infrastructure are not accounted for here). In particular, the geo-location is always assumed feasible, since a problem exposing itself as a high CVSS scoring that always concerns a known component, which in turn has a known geographic location. Figure 5 shows an example of a high-level perspective on a large-scale enterprise infrastructure, where mostly the subsidiaries are being displayed. In refining the picture to show subsystems, it is a simple matter to colour the nodes according to the (aggregated) risk scores, to dig into the "why" of a certain risk situation. An example is displayed in Figure 6.

4.1.2 Deriving contingency plans

With the process being automated to the described extent, it offers a good support for making plans towards lowering unacceptably high risks. If the high level view displays a high risk, then zooming into the picture will reveal the origins of the higher risk. Subsequently asking for geographic references or pointers within the physical network topology aids the security officer in identifying places where to apply additional security.

Assessing the effect of additional security is then a matter of re-doing the security assessment but under changed conditions. That is, the security officer may select certain additional precautions and then needs to define how these would change the CVSS scoring. Given an automated CVSS survey sheet that computes the necessary values upon entry (like in a spreadsheet program), the output values can be fed back into the risk

aggregation mechanism to update the risk picture. The rest of the process is then iterative refining and trying different fixes until a satisfactory solution is found. Technically, this is a matter of combinatorial optimization, which may as well be automated, provided that there aren't too many options to try.

Figure 5 Refinement of risk visualization (König, et al. 2015)(König, et al. 2015)

Figure 6: Color codes for different risk levels (König, et al. 2015)

5. Conclusion

In this work, we proposed a twofold add-on to standardized risk assessment methods, which consists of an a-priori questionnaire for experts to reduce answer variability, as well as an automated aggregation method based on neural networks. Generally, the problem of risk assessment and aggregation is still widely up to manual labour and human expertise. Since opinions about risk may be different and occasionally disagreeing, the here proposed meta metric (section 3.1.1) shall help in making a "more informed" decision by asking relevant questions beforehand to guide the risk expert. It has been empirically observed that divergences in the answers are reduced by this approach, and refining, adapting and designing scenario-specific such meta-metrics appears to be a promising part of future work. A statistical significance of the meta-metric in terms of a reduction of variance could, however, not be tested due to insufficient response data (which in turn is difficult to obtain due to the observed scarcity of expert knowledge related to CVSS). A first step of future work is thus the statistical confirmation of the observations made herein, based on a larger sample.

Secondly, the use of neural networks offers the particular appeal of letting a risk expert "give away" her/his expertise on risk aggregation. To this end, we trained the network with aggregation that experts did manually (in interviews), and validated the network's functionality on the complement subset of the expert data (as only part of it was used for training). The overall risk scoring is obtained in a bottom up direction starting from individual components up to the top of the entire system. The "top-scoring" can then be broken up into its details as demanded (thus allowing the decision-maker to "zoom-in"), and can additionally be enriched by other infrastructure-related information to refine the picture if needed. Among such needs is, e.g., geographic referencing, filtering, modularization or views on different levels (organizational, technical, etc.). Such visualizations are helpful in Security Operating Centres (Cyber-War Situation rooms), where the current situation is available in real-time and newly evolving threats (arising from changes to the system or recently discovered exploits) can be countered properly and timely. Our proposed risk aggregation and visualization technique supports all these matters and is tailored to large scale applications where the risk situation is infeasible to analyse on the level of individual components. Our findings show that aggregation and visualization towards a concise yet informative risk picture is effectively and efficiently possible.

Finally, exploiting the usual tree-like/hierarchical structure of many networks in connection with an automated risk aggregation offers, we can aid a risk manager in finding resolutions for problematic situations, simply by

letting the system re-calculate the risk in different scenarios. Especially in times where security incidents must be reacted upon very timely, such tool aid appears more than necessary.

References

Alpcan, T. & Başar, T. Network Security: A Decision and Game Theoretic Approach Cambridge University Press, 2010.

Beck, Alexander; Trojahn, Matthias and Ortmeier, Frank. "Security Risk Assessment Framework." *D-A-CH Security*, 09 2013.

Beck, Alexander. "Aufbau eines neuronalen Netzes zur Schwachstellenaggregation." *Datenschutz und Datensicherheit Journal*, 06 11 2014.

Bundesamt für Sicherheit in der Informationstechnik. „IT-Grundschutz-Kataloge". Bonn, Deutschland, 2013.

Heidorn, Alexander. „Prototypische Implementierung eines Security Risk Assessment Frameworks (SRAF) zur Erstellung und Aggregation von SRAF Graphen". Wernigerode: Hochschule Harz - BA, 2014.

Jajodia, Sushil; Noel, Steven; Kalapa, Pramod; Albanese, Massimiliano; Williams, John: "Cauldron: Mission-Centric Cyber Situational Awareness with Defense in Depth," *30th Military Communications Conference* (MILCOM), Baltimore, Maryland, November 2011.

Jansen, Wayne. In *Directions in Security Metrics Reseach*, 9-14. Computer Security Division, NIST, 2009.

Joh, H.; Malaiya, Y. K.: „Defining and Assessing Quantitative Security Risk Measures Using Vulnerability Lifecycle and CVSS Metrics," *Int. Conference on Security and Management (SAM11),* pp. 10-16, 2011.

Kawasaki, Ritsuko and Hiromatsu, Takeshi: "Proposal of a Model Supporting Decision-Making on Information Security Risk Treatment", *International Journal of Economics and Management Engineering* Vol:1, No:4, 2014.

König, Sandra, Stefan Rass, Stefan Schauer, and Alexander Beck. "Risk Propagation Analysis and Visualization using Percolation Theory." *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7 1 2015.

Lenges, Michael. "Framework zum IT-Risikomanagement: Integriertes Betriebssicherheitsmanagement der Geschäftsprozesse und Datenverfügbarkeit". Books on Demand, 2009.

Liu, Fang; Dai, Kui; Wang, Zhiying and Ma, Jun: "Research on Fuzzy Group Decision Making in Security Risk Assessment", in P. Lorenz and P. Dini (Eds.): *ICN 2005*, Springer LNCS 3421, pp. 1114-1121, 2005.

Mell, Peter, Karen Scarfone, and Sasha Romanosky. "A Complete Guide to the Common Vulnerability Scoring System". First.org. 2015. https://www.first.org/cvss/cvss-v2-guide.pdf (accessed 12 11, 2015).

Miura-Ko, R. Ann; Yolken, Benjamin; Mitchell, John and Bambos, Nicholas: "Security decision-making among interdependent organizations", in Proc. of 21st IEEE Computer Security Foundations Symposium, IEEE 2008.

National Institute of Standards and Technology (NIST): "Managing Information Security Risk", Special Publication 800-39, March 2011

Nelsen, Roger B. "An Introduction To Copulas". Springer, 1999.

Noel, Steven; Harley, Eric; Tam, Kam Him and Gyor, Greg: "Big-Data Architecture for Cyber Attack Graphs: Representing Security Relationships in NoSQL Graph Databases," *IEEE Symposium on Technologies for Homeland Security (HST)*, Boston, Massachusetts, April, 2015.

O'Hare, Scott; Noel, Steven and Prole, Kenneth: "A Graph-Theoretic Visualization Approach to Network Risk Analysis", in J.R. Goodall, G. Conti, and K.-L. Ma (Eds.): *VizSec 2008*, Springer LNCS 5210, pp. 60-67, 2008

Rass, Stefan, and Sebastian Kurowski. On Bayesian Trust and Risk Forecasting for Compound Systems. *Proceedings of the 7th International Conference on IT Security Incident Management & IT Forensics* (IMF), IEEE Computer Society, 2013.

Rass, Stefan: "On Game-Theoretic Risk Management", arXiv:1511.08591 and arXiv:1506.07368.

Röcher, Dror-John. "Metrikbasiertes Patchen mit CVSS 2.0 Konzept mit Methode." *ERNW Newsletter*, 09 2007.

Xinlan, Zhang; Zhifang, Huang; Guangfu, Wei and Xin, Zhang: „Information Security Risk Assessment Methodology Research: Group Decision Making and Analytic Hierarchy Process", in *2010 Second WRI World Congress on Software Engineering*, IEEE Computer Society, 2010.

Zell, Andreas. Simulation neuronaler Netze. Oldenbourg, 1997.

A Method to Generate SQL Queries Filtering Rules in SIEM Systems

Martin Dvorak
Department of System Analysis, Faculty of Informatics and Statistics, University of Economics, Czech Republic
dvorakmar@seznam.cz

Abstract: Organizations, within their risk management, implement SIEM systems (Security Information and Event Management) to ensure information security. These systems evaluate data from various applications used within an organization's IT environment, and decide whether a security incident happened or not. The problem is that these applications generate too much data; this makes the SIEM system's operation costly both in terms of license fees and infrastructure requirements. The author defines a new method to optimize the amount of data entering the SIEM system while maintaining the level of an organization's information security. The method aims to reduce the SQL queries that are sent to the system for evaluation. The aim of this article is to introduce this new method for reducing data processed in the systems of evaluating information security events.

Keywords: information security, cyber security, SQL queries filtering, event security evaluation, SIEM

1. Introduction

Organizations have been carrying out an increasing number of activities electronically, with a growing proportion of activities which can be seen, in terms of information security, as risky ones, see – (CZECH STATISTICAL OFFICE, 2015). At the same time, a number of security incidents focused on financial gain or theft of sensitive information have been continuously growing up. The assailants have carried out more sophisticated attacks not only with the help of information technology, but also through social engineering methods. Government institutions, aware of this growing trend, adopt various measures to help enhance information security in general. The recently adopted "*Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union*" can be mentioned as evidence of this effort.

It is not only the regulation of cyber security that requires monitoring of events in order to achieve – by an organization defined – a level of organizational information security. Monitoring and assessment of safety occurrences is in accordance with best-practices and important standards, such as the international standard ISO / IEC 27001, which defines information security management system. Implementation of systems that evaluate events is relevant to the organizations not only in terms of meeting legal and regulatory requirements, but also in terms of implementation of measures against the risks that organizations face. According to the Verizon global survey (VERIZON, 2015), which examines the causes of safety incidents in 2015, one can implement a SIEM system to cover approx. 64,9% of security incidents (see Figure 1); i.e. specifically, against the crimeware, cyber-espionage, insider misuse, web applications attacks and errors.

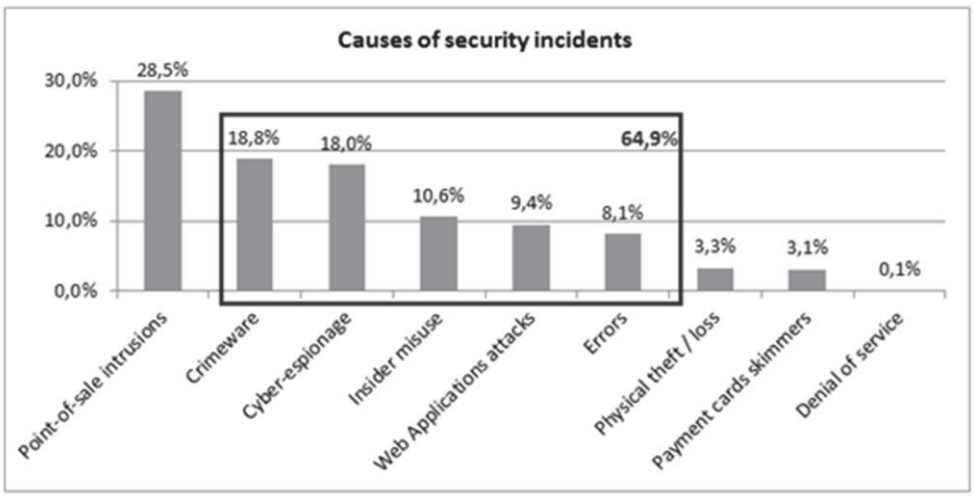


Figure 1: Causes of security incidents. Source: (VERIZON, 2015)

1.1 Problem definition and article objectives

The subject of this research work is an organization's information security (hereinafter IS), reached by a system evaluating information security events (hereinafter SIEM). SIEM systems evaluate the data from the applications within an organization's IT environment, and – based on that – decide whether there is a security incident or not. The problem is that the applications generate too much data / too many events which the system must evaluate. This makes the SIEM system's operation costly both in terms of direct and indirect costs. Direct costs relate to high license fees, because a majority of SIEM solution providers apply a license model based on the amount of processed data / events, for example (SPLUNK, 2015); indirect costs result from the amount of data processed by the SIEM system. The more data the system must evaluate, the higher are the demands on the computing power of the system, as well as demands on the network infrastructure[1] of an organization, which must transport the data into the system for processing.

Systems evaluating the events in terms of information security work in real time. The problem researched in this article is that within the critical infrastructure, the applications generate too much data and not all of this data is relevant to information security. If all the data were entered into the system evaluating events, its operational performance might be limited; this would be manifested by the system slowing down, resulting from its supersaturation or in complete congestion and subsequent malfunction of the system. As a result, the system would not be able to assess events in real time, which represents a key feature of these systems; SIEM would need more time to evaluate events which would cause delays in the user's work and thus reduce his/her comfort.

The aim of the research presented here is to propose a method which allows reduction of the amount of data entering the SIEM systems for evaluation. This method also allows the organization to optimize the operation of SIEM.

1.2 Related articles

In this work, the author is building on his previous research in which a normative framework of an organization's IS and principles of IS audits were defined (DVORAK, 2011), as well as on the work (DVORAK, 2013), in which the preconditions for the proper functioning of systems evaluating information technology (IT) users' behavior were defined.

Within the scientific community, several approaches leading to optimization (reduction) of SIEM systems' operational costs have been explored. Efforts to optimize SIEM systems' operation can be observed both on the input and output sides of SIEM system. Optimization on the input side aims to reduce the amount of entering data; the method proposed and discussed by the author of this article builds on this principle. Optimization on the output side aims to reduce the number of alerts the SIEM system produces, which Security Operation Center (SOC)[2] employees and system administrators would otherwise have to pay attention to. Another approach to optimizing SIEM system work is to enhance the internal logic of the system, which evaluates the events according to defined rules.

The author (HORNE, 2015) presents the concept of packet filtering on the SIEM system input side, in the context of detecting malware. He works with modelling traffic on the network and differentiating it from "known", which is filtered out, and unknown, which is dropped for further evaluation to the SIEM system. The disadvantage of this approach is that it does not cover a broader spectrum of IT risks arising, for example, from the human factor (error, data theft, etc.) presence, because it operates at a too low level of ISO / OSI model.

In their work (SUAREZ-TANGIL, 2015), the authors apply findings from the area of artificial intelligence and present a correlation module[3] based on genetic programming and neural networks, which is able to design and learn new correlation rules. This approach, which is an example of improving the internal logic of the work SIEM, greatly facilitates the work of the SOC operators in defining new rules, assessment of security events. On the other hand it does not reduce the data flow which SIEM system must evaluate.

[1] Requirements on the bandwidth of a communication channel.

[2] Team ensuring activities related to information security management.

[3] "Correlation Engine".

A method focusing on output from the SIEM system optimization is presented in the article (PECCHIA, 2014). This method builds on a filtering of alerts coming from a SIEM system. This approach reduces the necessary amount of SOC staff (FTE)[4], as the staff can focus only on the relevant messages.

Other scientific article (CATES, 2015) deals with IS development in recent years, its basic principles and trends.

The aforementioned approaches, including the method presented in this article, are not mutually exclusive, so organizations can – in order to achieve optimal SIEM system operation – choose several approaches according to their internal IT environment and architecture.

1.3 Definition of basic terms

The objective of this subchapter is to define the most used terms to achieve a unified understanding of the studied subject.

Information security

The following definition of the information security concept has been applied in the scientific community: information security means keeping the confidentiality, integrity and availability of information, as well as of other features such as authenticity, accountability, non-repudiation, and reliability. This definition has been integrated into the international standard ISO / IEC 27001.

Event

For the purpose of this article, under the term event will be understood any action in the organization IT environment that: a) is so important for the organization information security, it is worth noting it; b) is time bounded; c) additional attributes - namely where and when it took place, the subject of the event, event type (reading, editing, deleting) and who initiated it – can be unambiguously assigned to it.

Security Incident

For the purposes of this article, under the term security incident will be understood an event that involves a breach of information security (availability, integrity and confidentiality). A security incident can occur without an apparent detriment at the moment of the incident occurrence. The detriment can manifest itself later.

2. Basic principles of events assessment

In this section the basic principles of SIEM systems' functioning are defined. The method of data reduction defined below builds on these principles. The author compiled these principles as a synthesis of findings made during an analysis of the existing SIEM solutions. The analysis worked with the three best rated products (IBM QRadar, HP ArcSight, Splunk), according to consulting firm Gartner (Gartner, 2016).

Events' evaluation systems work as follows: using probes, these systems monitor data and events within an organization's IT environment; then – using statistical methods (correlation) – the systems create chains of events that have one common denominator: usually the user. The chains of events and activities thus represent a single-user's activities organized in a time sequence as initiated by the user. In order to assess if the particular chain of events represents a so-called critical chain of events of security incidents, the system needs to know the context in which the events occurred. This context is compiled on the basis of additional data from inside and outside of the organization. In other words, a system – based on data collected from data sources – can answer who, when, where and why performed what: what user initiated or from what IP addresses was the event initiated; what was the content of the event (data editing, copying, deleting); at what time did it happen; in which application did the event occur, and why did the event occur.

A SIEM system evaluates cyber events based on defined rules. Most systems have characteristics of "learning" which means that when in operation they are able to create new rules to evaluate events. However, the initial

[4] Full-time equivalent (FTE) is a unit applied to express requirements on the amount of work necessary to do working task.

evaluation rules must usually be defined by the SIEM system administrator based on the organization's security policy.

Based on these rules, each event is assessed and a decision is made into which category of events it belongs: a known safe condition; status unknown; a known unsafe condition. For each category, subcategories can be defined within the SIEM system to which one can assign a specific set of operations that the SIEM is to take if they occur. For example, when an unknown state event occurs, SIEM sends an entire event for further analysis to the SOC team members – then they may choose to create a new rule that defines how to evaluate such an event; in the case of the known unsafe condition, a SIEM may generate an alert or stop the activity completely.

The figure 2 below shows a very simplified diagram of the organizational IT architecture, including location of probes that deliver data to the SIEM system for evaluation. DAM[5] tools, which collect event data in databases, are used to connect a SIEM system to databases. Data flow reduction resulting from data filtration occurs in communication between the DAM and SIEM.

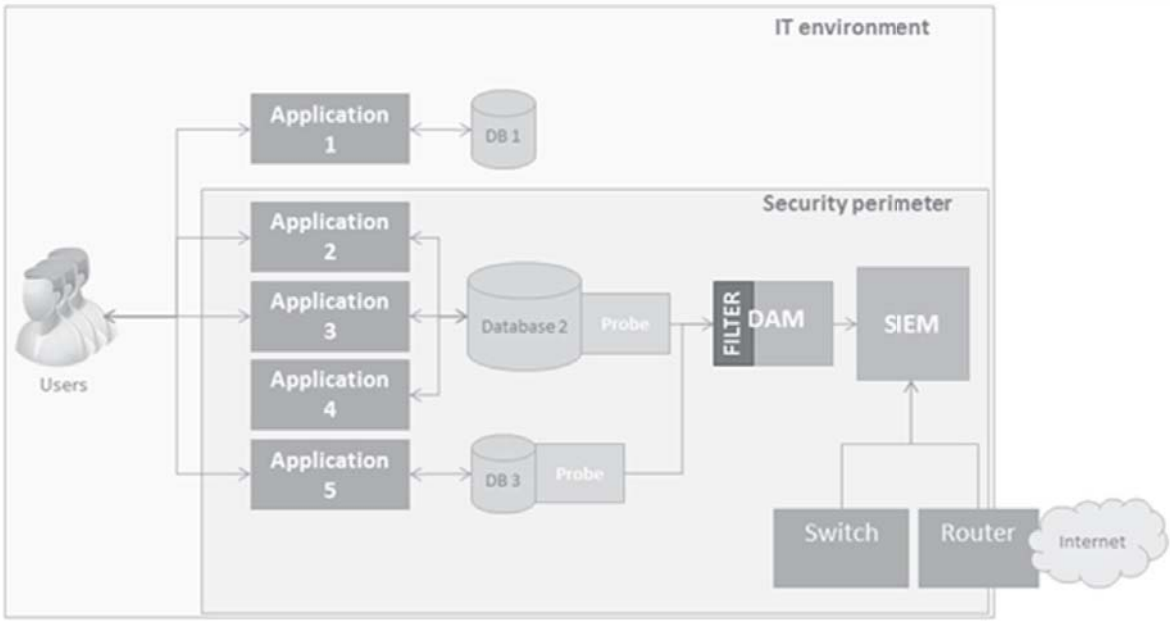


Figure 2: Simplified diagram of the data collection to SIEM system (source: author)

Optimization of a SIEM system operation almost always brings/means reduction of data that enter it; thus it is always necessary to find a compromise between a sufficient level of organization's IS and a sufficient amount of relevant data. In the case of an excessive restriction of data which are to enter the system, it might not be able to evaluate the events. That ultimately may result into an inability of the system to identify a security incident, or may generate too many alerts.

3. Data optimization

To meet the above-defined objective of the research, it is necessary to limit the amount of data entering the system evaluating information security events. For this purpose, the author proposes a method as described below. This method has the following assumptions:

- The organization has implemented an information security management system, for example, ISO / IEC 27001. The reason for this is that under such conditions, an organization has a defined security perimeter, classified data, and information security management policy.
- The organization plans to implement or has already implemented a SIEM system, separate from the information system; communication with databases is via probe (DAM).

The information system itself works with data from a database. This assumption, however, is not essential to the functionality of the method; the method can be reformulated for other models of the organization's IT architecture.

[5]The Database Activity Monitoring system serves as a connector, which enables to connect SIEM and to monitor activites in databases.

The central idea of the proposed method lies in the implementation of the filter layer between the probe and SIEM. For effective operation of the filter layer, it is required to identify sensitive data for each probe. The actual implementation of the filter is dependent on the used data base. Assuming a database platform, to implement such a filter requires designing a parser of SQL queries that decides whether the queries are dealing with sensitive data or not.

The method to optimize the amount of data entering the system evaluating events includes the following steps:

- 1. Identification of sensitive data
- 2. Identification of applications processing sensitive data
- 3. Compilation of data filtering rules
- 4. Implementation of data filtering
- 5. Verification and possible adjustment of filtering rules

Steps 3-5 can be repeated iteratively until the optimal amount of data that enters the SIEM system is reached. A description of each step follows.

3.1 Identification of sensitive data

Sensitive data is identified in accordance with the definition of a security perimeter. We can distinguish two groups of sensitive data:

- Identifiers (table columns), which are sensitive themselves. In this group is included each identifier, which is within the safety perimeter marked as sensitive data. The set of all identifiers that are sensitive in themselves, will be labeled I_c.
- A set of identifiers, which together constitute sensitive data. This group includes such sets of identifiers that are within the security perimeter identified as sensitive, i.e. those sets of identifiers that make it possible together to fill the role of a sensitive identifier (see point 1). The collection of all such sets of identifiers will be labeled I_m.

Example:

A birth number is always sensitive data because it unambiguously identifies a person and allows one to access personal data of a specific entity.

A surname (in a similar situation) itself is not sensitive data, as it does not unambiguously identify a person. In terms of information security, a surname can be seen as an unattractive identifier in this case. However, if the surname is combined with other data such as date of birth (which is also apparently uninteresting by itself), we get a set of identifiers that can define a sufficiently small group of people, and thus data obtained through a combination of identifiers can be considered as sensitive.

It should be noted that a set I_m is not defined unambiguously (see 2). For example, the set of identifiers $I_1 = \{\text{name}, \text{surname}, \text{date of birth}\}$ will probably be a desirable component of a set I_m, as well as will be $I_2 = \{name, \text{surname}, \text{address}\}$ and $I_3 = \{\text{address}, \text{date of birth}\}$. Apparently $I_m^1 = \{I_1, I_2, I_3\}$ will comply too, but it will also probably comply $I_m^2 = \{\{\text{name}, \text{surname}\}, I_3\}$, and if $I_m^3 = \{I_1, I_2, I_3, \{\text{name}\} \cup I_3\}$. A set I_m^2 sets less strict filtering: indeed it regards as sensitive data also those data that by definition of a security perimeter may not be sensitive. A set I_m^3 includes a redundant set of identifiers$(\{\text{jmeno}\} \cup I_3)$. Any even related to this set of identifiers, obviously will be captured within a set of identifiers I_3.

Therefore, the most effective filter will be the one based on a set I_m^1. However, the minimal set containing all the necessary sets of identifiers may not generally be easy to find. If the data platform includes n identifiers, there are 2^n sets of potential identifiers. Thus, finding the optimal set I_m is not generally possible in reasonable (polynomial) time. The proposed method partially eliminates this imperfection within steps 3 to 5 via a heuristic process/improvement.

The output of this step is therefore represented by sets of I_c identifiers of sensitive data (data items that contain sensitive data) and by I_m of sets of identifiers, which together lead to sensitive data. For simplicity of a following explanation let's assume that I_c are single element sets.

3.2 Identification of applications processing sensitive data

The objective of this process step is to identify in which applications and how sensitive data is processed. The first part of analysis (3.1.) basically consists in identifying sensitive data flows through the organization applications; the second part (3.2.) of the analysis is primarily concerned with finding the way by which data items are sought, which data items can appear in these applications, and what options users have for reading, editing or deleting specific data items.

Analyzed are always those applications that fall within the defined security perimeter based on the organization's security policy. However, a scheme of architecture or data model of the organization can be used, if available.

For each set of sensitive identifiers I, for each application in the security perimeter, it is registered whether it is necessary to monitor queries in the filter layer.

If we denote A the set of applications falling within the security perimeter, the function $f: I_c \cup I_m \times A \mapsto \{0,1\}$ is an output of this step. If I is a set of identifiers and a is an application, then it is necessary to monitor the set of identifiers I in application $f(I,a) = 1$, otherwise $f(I,a) = 0 f(I,a) = 0$.

If it is for organizational reasons desirable to distinguish between actions (read, edit, delete) that can be in an application a carried out above a set of identifiers I, it is possible to introduce a function f as $f: I_c \cup I_m \times A \mapsto 2^O$, where $O = \{\text{delete}, \text{reading}, \text{editing}\}$ and 2^O is a set of all subsets of the set O.

Optionally, for clarity, the functional values of the function f can be stored in the matrix, where the lines correspond to single sets of identifiers and columns to different applications.

Table 1: Example matrix with set of identifiers per each application. source: author

Identifier	Application			
	SAP	Siebel	EXK	…
Social Security Number	X	X	X	…
ID of a person in database	X	X	X	…
Surname + Date of birth		X		…
Surname + Address	X	X	X	…
Address + Date of birth		X		…
…	…	…	…	…

3.3 Compilation of data filtering rules

For each database (each probe), it is required to implement those filtering rules that will test whether each individual event in application initiated inquiries (SQL queries) over the sensitive data.

Similar to the construction of sets of identifiers comprising the sensitive data, the calculation may encounter the limits of reasonable time requirements. Indeed, it is easy to recognize that the test of whether a specified query asks for one of the sets of identifiers compiled in step 1, it is necessary to convert the SQL query to disjunctive normal form (hereinafter DNF), or perform an equivalent operation. Transfer to DNF is generally NP-difficult. It is therefore necessary to seek a reasonable heuristic method.

The proposed method circumvents the problem of DNF as follows: a number p of minterms (literals associated by conjunction) that is admissible to be constructed is determined. If during the DNF construction it gets evident that the DNF contains more minterms than p, the query is automatically designated as sensitive.

The number p can be optimized in steps 3-5.

If the DNF can be constructed, it is further checked whether any combination of identifiers from the set $I_c \cup I_m$ of identifiers is a subset of identifiers in any minterm. If so, we consider the query as sensitive, otherwise it can be filtered out.

The test described above can be realized (by "brute force") in time $O(np|I_c \cup I_m|)$, where n is the number of identifiers and $|I_c \cup I_m|$ indicates the number of elements of the set $I_c \cup I_m$.

3.4 Implementation of data filtering

The aim of this step is to create a functional prototype of a filter in the form of either an individual application or a module of a database probe of the SIEM system.

The functionality of the created filter consists in a verification if the defined filtering rules are met; then when the rules are met, SQL queries are sent for further evaluation in the SIEM system. If the rule is not met, the filter will not let an SQL query enter the SIEM system.

The outcome of this process step is an application / module that filters SQL queries, so only the SQL queries that are relevant to the IS organization enter the SIEM system.

3.5 Verification and possible adjustment of filtering rules

The constructed filtering rules or the aforementioned array of sensitive data from process step 2 (if constructed) can be used for verification purposes. Both of these serve as a test scenario. The tester sends SQL queries according to the matrix and then verifies that the SQL queries – if necessary – were or were not reflected in the SIEM system. If the results match the defined filter rules, the application can be put into actual operation.

The actual verification of the proposed method is described below.

3.6 Key roles and responsibilities in the process

Based on the experience from the implementation of this method in practice, author compiled a matrix of key roles and relevant responsibilities in the data reduction process – a so-called RACI matrix (Responsible, Accountable, Consulted, and Informed). This matrix is described in the table 2. For completeness, a key input and output necessary to implement the relevant step is added to each step.

Table 2: Definitions of key responsibilities and roles in the process. source: author

Definitions of key roles and responsibilities in the process							
Step	Input	Output	Security Officer	Application Specialist	Database Specialist	Programmer	Tester
Identification of sensitive data	Organization security policy Classification of data in an organization Data model of an organization	Matrix of sensitive data	A	C	R		
Identification of applications processing sensitive data	Matrix of sensitive data Scope of information security management in an organization Scheme of application architecture Organization data model	Updated matrix of sensitive data	A	R	C	I	I
Compilation of data filtering rules	Matrix of sensitive data	Data filtering rules	A	C	R	C	I

Definitions of key roles and responsibilities in the process							
Implementation of data filtering	Data filtering rules Matrix of sensitive data	Module filtering SQL queries	A	C	C	R	I
Verification and possible adjustment of filtering rules	Matrix of sensitive data Data filtering rules	Confirmation of compliance of filtering rules with reality	A	I	I	C	R

4. Verification of the proposed method

The author verified the above proposed method empirically via a practical application in an organization processing personal data. For security reasons, the organization does not want to be named. The organization has a three-tier IT architecture that respects SOA[6] principles, a central data repository, and has approximately 8,000 employees. The above proposed method led in the described organization to creation of 249 filtering rules. The verification was done on embedded IBM QRadar system (and IBM Guardium system serving as DAM tool) in two steps:

- 1. Verifying if the proposed method helps to reduce the amount of data entering the SIEM by measurement of the number of SQL queries before and after data filtering.
- 2. Evaluating whether the method does not reduce the level of organization's IS because, as a result of the method application, the SIEM system does not receive all data on SQL queries.

In the first step of verification, one metric was monitored – the number of SQL queries. The verification process took 7 weeks, from November 2nd to December 18th, 2015. Of the collected sample, weekends and public holiday (17th November, 2015) were excluded. Measurement concerned all applications processing sensitive data[7]. Measurements were carried out parallely on the database input before applying filtering rules, and the output of the probe after application filtering rules (the SQL queries that are sent to the SIEM system for evaluating). The advantage of parallel measurements at two points at the same time is that seasonal and other interferences are eliminated.

In front of the filter module, an average of 2 439 184 SQL queries per day was measured, which would come into SIEM for evaluation. After application of the proposed procedure for data optimization, the amount of queries decreased to average 1 439 333 SQL queries per day. The average number of SQL queries after the application of the proposed procedure for data reduction is lower by 43%. Detailed average numbers of queries for each day of the week for the period of measurement is shown in the figure 3. It can be concluded that the proposed method significantly reduced the amount of data that must be evaluated by the SIEM system.

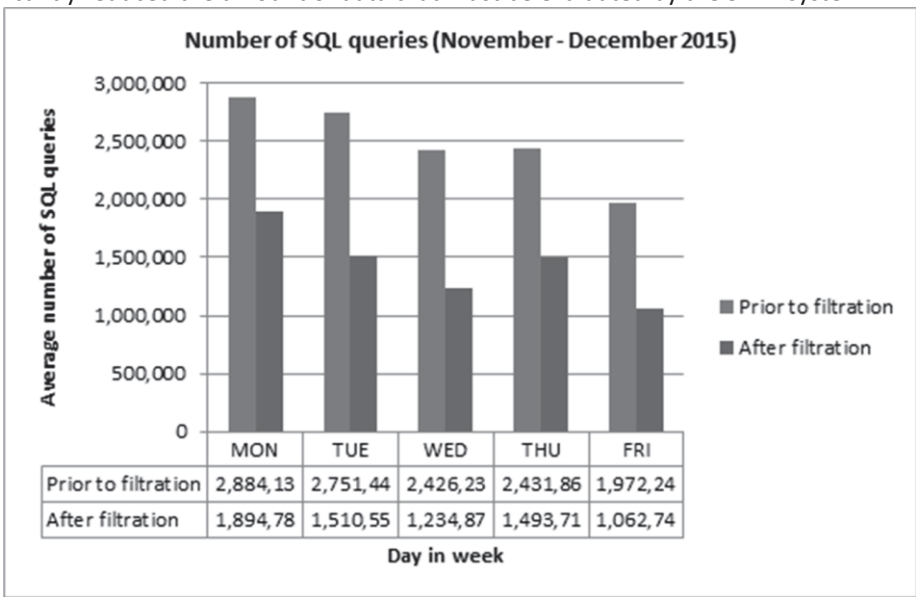


Figure 3: The number of SQL queries prior and after application of filtration rules (source: author)

6 Service Oriented Architecture

7 All applications within a security perimeter of the organization.

Queries filtered out can be divided into several categories. These categories are sorted according to their proportion in the group:

- 1. So-called "Dummy" queries or "dummy selects" (34,05%); for example, the types of queries like these: *Select "1" from Dual* or *Select sysdate () from dual*. This type of query can be used by the application to validate the database connection, but in terms of loss of sensitive data, is irrelevant. Into this group also fit queries like store log entries: "*Insert into log ...*"
- 2. Queries on the enumerators (23,54%) – these queries typically return values of enumerators, so it can be considered from a security perspective as irrelevant.
- 3. Procedures on the database: *proc delete_log (day)* (22,6%). Some applications can call predefined procedures that the database is to perform. Procedures working with sensitive data must be identified in step 2, and subsequently filtering rules need to be modified. The procedure mentioned in the example deletes a log file for that particular day.
- 4. Into the last category fit those queries which were not aimed to obtain sensitive data (19,81%). These are such queries that after parsing on a so-called SQL construct (which consists of a command, object, and field) do not work with predefined fields of table that contain sensitive data.

When comparing the two graphs before and after deployment of filter rules, it is possible to notice a higher variability in the number of SQL queries during the reporting period. This higher variability was also analyzed.

Transportation of SQL queries from applications to the database has several components:

- 1. "Fixed transport" – this transport appears every day and is pretty unrelated to the activity of users in applications. For example, applications provide daily statements, count reports, etc.
- 2. SQL queries from neighboring systems – for example, input from the "real world" in the form of a daily dose of imports from data boxes.
- 3. SQL queries caused by the intrinsic activity of users.

The second step of the verification lies in an assessment whether the method does not contribute to a reduction of the organization's IS because it reduces the amount of data entering SIEM system. For that purpose, the author performed a comparison of reports on the average number of security incidents in a week before and after data filtering. Relevant measurements were carried out sequentially: measurement before filtration of data implementation was carried out for the months of February-March 2015; measurement after data filtration implementation was run in the period from April to May 2015. By comparing the two reports, it was found that the average number of security incidents per week remained at the level of a few incidents per week both before and after filtering, which means that there were no significant differences.

As a third step of the proposed verification process, the author planned to examine an amount of alerts from the SIEM system before and after filtering the data; however, the organization refused to grant them permission to publish the obtained data.

Benefits for the organization which have applied the above proposed method were: reduced requirements on the bandwidth of the communication channel, storage capacity and reduced requirements on processing power of the servers by 30%.

Based on the results of verification acquired during the two steps described above, it can be concluded that the proposed method significantly reduces the amount of data entering the SIEM system for evaluation without reducing the IS of an organization; thus, the objective of the research presented in this paper was reached.

5. Conclusion

The aim of the research presented in this paper was to design and validate a new method that reduces the amount of data processed by a SIEM system. The method, proposed by author, is based on the analysis of sensitive data, and on the way in which this data is processed within an organization’s IT environment, including the subsequent application of logic functions enabling to define filtering rules that restrict the amount of data entering the SIEM system.

This method was verified empirically by measuring the number of SQL queries and security incidents. These measurements confirmed that the proposed method makes it possible to significantly reduce the number of events (SQL queries) that otherwise would have to be sent to the SIEM system for evaluation, all that without compromising the level of the organization's information security. In our case, a surprisingly large reduction was achieved (43% of average number of SQL queries). This can result, besides other things, from a low level of application software used in the searched organization, which generates a large amount of – from our perspective – ballast SQL commands.

During the verification process of the proposed method, the author also identified other possible improvements in the optimization process, reached through SQL queries' monitoring and filtering based on a search of the query initiator. An implementation of metrics for individual search data that could enable to carry single SQL queries' scoring – similar to determining whether an e-mail message is a SPAM – looks promising. The author wants to check this hypothesis in their following research.

References

Cates, S. (2015) *The Evolution of Security Intelligence*. Journal Network Security. 2015, Elsevier Science Publishers B. V. Amsterdam, The Netherlands.

Czech Statistical Office. (2015) *Informacni technologie v podnikatelskem sektoru*. Available online: <https://www.czso.cz/csu/czso/podnikatelsky_sektor>

Dvorak, M., Rihova, Z. (2011) *Problems of ISO 27001 Matrix certification*. Prague 25.05.2011 – 26.05.2011. In: Information Security Summit. Praha : TATE International, 2011.

Dvorak, M. (2013) The preconditions for the management of IT security based on standard user behaviour.

ECON. Vol. 23. 2013

Gartner. (2016) *Magic Quadrant for SIEM*. Available online: <https://securityintelligence.com/ibm-is-a-leader-again-in-2015-gartner-magic-quadrant-for-siem/>

Horne, W. (2015) *Collecting, Analyzing and Responding to Enterprise Scale DNS Events*. In: CODASPY '15 Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. ACM New York, 2015.

International Organization For Standardization. (2014) *ISO/IEC 27001 Information technology - Security techniques - Information security management systems - Requirements*.

Pecchia, A. Cotroneo, D., Ganesan, R., Sarkar, S. (2014) *Filtering Security Alerts for the Analysis of a Production SaaS Cloud*. In: UCC '14 Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. IEEE Computer Society Washington, 2014.

Splunk. (2015) *Splunk pricing*. Available online: <http://www.splunk.com/en_us/products/pricing.html>

Suarez-Tangil, G., Palomar, E. And Rigaborga, A. Sanz, I. (2015) *Providing SIEM systems with self-adaptation*. Information Fusion Journal. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, 2015.

Verizon, (2015) *2015 Data breach investigations report*. Available online: <http://www.verizonenterprise.com/DBIR/2015/>

Designing Real-Time Anomaly Intrusion Detection Through Artificial Immune Systems

Adriana-Cristina Enache and Valentin Sgârciu
University Politehnica of Bucharest, Faculty of Automatic Control and Computer Science, Bucharest, Romania
adryanaenache@gmail.com
vsgarciu@aii.pub.ro

Abstract: The sophisticated nature and large proliferation of cyber threats today requires efficient security solutions. In spite of the recent advances, security is still a challenging topic, as attacker's "skills" continually evolve with results seen in daily intrusion attempts. The human immune system defends us every day from "invaders". Therefore why not use this paradigm to detect intrusions? In this paper we propose an anomaly based IDS model with active response that combines two artificial immune systems (AIS) algorithms in a cascading manner: the Dendritic Cell Algorithm (DCA) (a second generation of AIS) and the Clonal Selection Classifier Algorithm (CSCA). DCA will offer online detection and by adding segmentation we succeed to improve the detection rate. We test our proposed IDS on the NSL-KDD dataset and examine the influence of segmentation for DCA. Furthermore, we compare the second component (CSCA) with other machine learning classifiers (SVM and C4.5).

Keywords: IDS, AIS, clonal selection, DCA

1. Introduction

Today security has been acknowledged as a priority for government and private sectors. However, the "number and frequency of publically disclosed data breaches is dramatically increasing" (Europol 2015). It seems that cybersecurity is deemed to a vicious circle where attackers are capable to outrun security solutions and even non-tech-savvy cybercriminals can acquire hacking tools from the Internet underworld and enter the cybercrime arena. Therefore we can deduce that as technology solutions evolve, they might also offer attack vectors and open doors for new types of cybercrimes.

Current threats "embrace" security mechanisms by encrypting user's data (the Cryptowall ransomware), authenticating or authorizing botmasters (Tinba Trojan), and target new technology ("dumb" smart devices, Internet of Things, Cloud Computing etc.) that may include simple vulnerabilities which can be escalated into complex cyberattacks. In this context, the security experts have gained forces and tried to mitigate these threats, including partially successful cyber campaigns (operation Tovar, operation Onymous etc.). Nonetheless, in order to address attacks generated by machines we need machines to countervail complex cyber threats.

Intrusion Detection Systems have become an omnipresent component of security solution architectures, mainly because they offer real-time detection, logging system events for further analysis and provide an additional line of defense (Enache & Sgarciu 2015). Based on the data analysis method there are two types of IDS, signature based and anomaly based. If the first approach identifies only known attacks, the latter can also identify new types of threats but it can also hinder the false alarm rate. In the current cyber medium where intrusions happen on a daily basis and new threats are identified, it is clear that in order to address intelligent threats we need intelligent security solutions capable to process large volumes of data in a real-time manner. Computational Intelligence (CI) can offer promising solutions to help mitigate cyber threats by using knowledge, continual learning and processing power.

Artificial Immune Systems (AIS), as a sub-field of CI, can add adaptability, flexibility, scalability and robustness to the IDS model (Kim 2002). Inspired by the principles and mechanisms of the biological immune system, it can help "secure" our system similar as the human immune system defends us from invaders. In this paper we combine, in a cascading manner, two Artificial Immune Systems to obtain a near real-time anomaly IDS distributed model. We implement the Dendritic Cell Algorithm for a primary and rapid detection, while for the final analysis we use the Clonal Selection Classifier Algorithm (CSCA). The rest of this paper is organized as follows: first we introduce AIS algorithms and show some related works, next we define the algorithms we use in our proposed model; in section 3 we describe our proposed IDS model, while in section 4 we analyze the test results obtained for the NSL-KDD dataset. At the end we show conclusions and state future works.

2. Artificial immune systems

The nature has always been a source of inspiration for researchers, as these systems have proven their efficiency in time and can offer optimal solutions while maintaining an equilibrium of the components involved. Although, the immune system is not completely known even by specialists in the immunology field, this has not stopped researchers to model the known paradigms into algorithms. Artificial Immune Systems (AIS) are bio-inspired CI algorithms, which transform the biological principles, functions and models of the immune system into mathematical models to help solve problems (de Castro & Timmis 2002). In general, the initiation of AIS is considered to have been in 1986, when Farmer et al. (1986) analysed theoretical aspects of immune networks and formalized them into a model. Work on AIS emerged in the mid 1990's, by works of Forrest et. al. (1994a) or Hunt & Cooke (1995), thus AIS became an area of its own.

Unlike other related algorithms, such as evolutionary algorithms, AIS has no archetypal model as its root, instead there are four major sub-filed that inspired AIS including: negative selection, clonal selection, immune networks and danger theory (Gu 2011). In the following subsections we show some related works and introduce the algorithms that stand at the basis of our proposed model.

2.1 Related work

There are many attributes that make AIS compelling for computer security, including IDS. First of all, these are CI algorithms and therefore carry with them their properties such as: using knowledge to solve problems, continual learning or efficient computational power in case of large datasets. Furthermore, AIS adds other benefits including (Kim et al. 2007): adaptability (it adjusts over time in changing conditions, which applies to the dynamic cyber context), robustness (it does not have a single failure point and if one component fails to perform its task the system will still operate), self-organization (there is no central command point) and scalability (required in the case of large data volumes needed to be processed).

One of the first AIS models to be proposed for computer security was introduced by Forrest et. al to detect computer viruses (Forrest et al. 1994b) and system call sequences (Forrest et al. 1996), which implied the negative selection paradigm, distinguishing between self and non-self components. Even though not many features of the immune system were used, authors took brave and important first steps in direction of AIS for information security.

The next significant contribution was the proposed IDS model based on a modified version of the previous negative selection, called LISYS (Lightweight Immune SYStem) (Hofmeyr & Forrest 1999). In this case the improved algorithm took into account the lymphocyte's life cycle and the proposed model was designed to detect malicious network connections in a distributed environment. The results of the model were promising and encouraged other researchers to pave their way into proposing other models based on LISYS.

StatiCS (Kim 2002) (Kim & Bentley 2001) is such an example based on a simplified version of LISYS, that uses clonal selection algorithm with negative selection operator in order to create a misuse detection model. The detector genotypes are represented as a binary string and a matching function is introduced. Tests on the UCI repository for machine learning showed good results, with negative selection maintaining a low false alarm rate. Authors also examine dynamic clonal selection (Kim & Bentley 2002) (Kim 2002), proving that it has learning and self-adaptation capabilities to novel data. Nonetheless, the two models have not been validated on larger networks.

With AIS at its early stage, we can observe that in the middle 1990's and early 2000's work on self non-self discrimination are predominant for information security problems, including simple Negative Selection Algorithm (NSA) or hybrid versions encompassing NSA. In spite of their early success, NSA does not scale well in large or high speed network connections, making the model to be inefficient because it requires a long time to generate a complete set of detectors.

Recent algorithms were inspired by the novel danger theory paradigm emitted by Polly Matzinger (1994), which claimed that the immune system does not distinguish between self and non-self, instead it responds to danger (called danger signals) that harms the host. Inspired by this theory, scientists have proposed the Dendritic Cell Algorithm(DCA) and applied it to anomaly detection problems such as host based detection of ping scans

(Greensmith et al. 2006), obtaining a 100% detection rate when threshold is appropriately set. Another application of DCA includes SYN scan detection (Greensmith & Aickelin 2007), which obtained a high detection rate and a low number of false alarms with experiments completed within minutes. Later on Botnet Detection (Al-Hammadi et al. 2008) in an IRC protocol based network is solved with DCA. Researchers showed that a single bot could be detected from the normal processes on the host.

Gu et. al. (2008) used the well-known KDD 99 dataset in order to test DCA for intrusion detection and compared it to C4.5 and a real-valued NSA. Results revealed that DCA is a promising candidate.

Given the promising results, researchers have tried to improve DCA. Chelly et. al. (2010) proposed a new classification method called fuzzy dendritic cell method(FDCM), with DCA in the Fuzzy Set Theory(FST) Framework. Mimicking human decisions, FST is used in this case to offer assessment for the immune algorithm. Results showed that FDCM smooths the separation between normal and abnormal, obtaining a better accuracy level than the original DCA. Kumari et al. (2012) combined DCA with Dempster-Belief theory to construct an intrusion detection model. In this case the probability of evidence that an antigen is an attack or normal is estimated by Dempster-Belief Theory. Tests on the KDD cup dataset, showed that the novel method could return higher accuracy and detection rate.

2.2 Dendritic cell algorithm

The Dendritic Cell Algorithm(DCA) is a second generation of AIS algorithms and it was inspired by the properties and mechanisms of dendritic cells that respond to some form of danger signals (Gu 2011). Proposed by Greensmith (2007), the algorithm is based on a population of identical cells that are at first immature and then become semi-mature or mature, depending on the type of antigens that it has encountered. The algorithm abstracts antigens as the items that will be classified, while signals are measures of status of the system. To evaluate the antigens, DCA combines them with signals and returns a Mature Context Antigen Value (MCAV) with a value between 0 and 1. To decide if an antigen is an anomaly or not, the MCAV is compared to a predefined threshold.

Each cell processes signals from the environment that influence their maturity state. Authors assume there are three types of signals including (Gu 2011):

- PAMP - the selected attributes of the signal indicate an anomaly; it has an important influence on the cell's transition from immaturity to maturity.
- Danger - the selected features most probably indicate an anomaly
- Safe - indicates a normal context.

In order assess signals and categorize them into three types, the algorithm involves a pre-processing stage. This can be done based on expert knowledge or as an automated process (Information Gain, Principal Component Analysis etc.). For our purposes we will use the IG method.

The cell has an encoded signal processing mechanism, that combines signals and weights into an output signal, and a lifespan dynamically chosen for each cell. The output signal (O_j) is calculated as antigens are sampled as follows:

$$O_j = \sum_{i=1}^{3} W_{i,j} S_j \qquad (1)$$

where S_j is the input signal (PAMP, Danger or Safe) and $W_{i,j}$ is the transforming weight from S_j to O_j ($j = \{1, 2, 3\}$) detailed in table I (Enache et al. 2015).

Table 1: Weight matrix

	PAMP	**Danger**	**Safe**
	S1	**S2**	**S3**
CMS (O1)	2	1	2
CMS (O2)	0	0	0
CMS (O3)	2	1	-2

As it is exposed to signals, the lifespan will be decreased with a value equal to its CMS output component. The cell reaches a maturity level when its lifespan has expired and in this stage the label of the antigen is memorised, while the cell is returned to the population into an immature state for further processing. This phase designates the detection stage.

Finally, the analysis is conducted by evidencing the signal profile of the cells for each antigen type and computing the value of MCAV as a proportion between the number of times the antigen was assumed anomalous and the number of times the antigen was exposed for analysis.

There are some variants of DCA, but the one described above is the deterministic one (dDCA). Furthermore, dDCA with antigen multiplication presumes that antigens are multiplied in order for them to be exposed to a larger number of cells and offer generalisation. This version is the one used in a former study (Enache et al. 2015). However, a drawback is that we need to wait for all the cells to finish their analysis of antigens, with multiplication. Thus, to offer online detection, in a near real-time fashion, we conduct antigen segmentation as described by Gu et al. (2009).

DCA has various advantages even when compared to other AIS algorithms, as it has fewer parameters and is an unsupervised classification algorithm. However, its performances depend on the selected threshold to be compared with MCAV, used in the analysis stage. A rightful concern is that its value might influence the performances of the algorithm. To address this issue we propose to apply a second "filtering" stage that will help improve DCA.

2.3 Clonal selection classification algorithm

Clonal Selection Algorithms (CSA) were inspired by Burnet's Clonal Selection Theory, which exploits the capabilities of B-lymphocytes to produce antibodies. Component of the adaptive immune system, clonal selection can be explained from the perspective of Darwin's theory as the evolution of antibodies in the immune system. The concepts that stand at the foundation of CSA include: antigens (candidate solutions), antibodies (detectors as a possible solution to the problem) and affinity (mechanism to evaluate antibodies). With this in mind, the main mechanisms of CSA are proliferation and mutation, which depend on the affinity level between antibodies and antigens (de Castro & Timmis 2002). While proliferation, or the number of generated clones, is directly proportional to the affinity level, the mutation will be inversely proportional to the affinity of the antibody and the antigen. Hence, the number of clones will be higher if the antibody matches the antigen and the mutation suffered by the clone will be less invasive should the affinity level be higher and vice-versa.

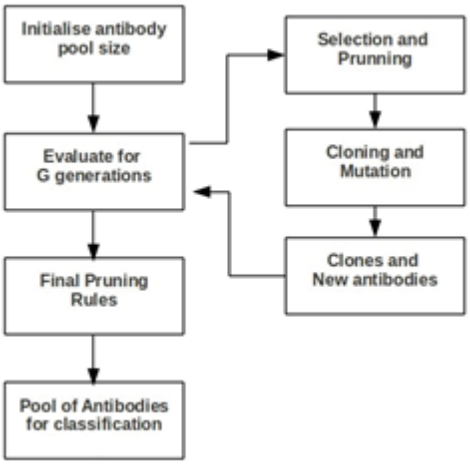


Figure 1: The clonal selection classification algorithm

There are some variations of clonal selection algorithms. Clonal Selection Classification Algorithm (CSCA) was proposed by BurnLee (2005) as a supervised classifier with the affinity mechanism, also called a fitness function, defined as an optimisation function that maximizes classification accuracy and minimizes misclassification. The training stage of CSCA consists of exposing several generations of antibodies to all the antigens from the dataset. Each generation is passed through (Brownlee 2005):

- *selection and pruning* -compute the affinity between the antibodies and antigens; based on the computed fitness value, apply selection rules to eliminate antibodies that have a misclassification score of zero, recompute the fitness value for antibodies with zero correct classification or a misclassification level of above zero, after class adjustment has been performed and finally eliminate antibodies that have a lower fitness value than a predefined constant.
- *cloning and mutation* - the selected set of antibodies is cloned and mutated according to their fitness value.
- *insertion* - the clones generated are included into the population of antibodies; finally a number of randomly selected antigens are added into the population.

Following the several tests of the generations, the resultant set of antibodies is exposed again to all the antigens, the affinity is computed and the final pruning rule is applied by removing those antibodies that have a lower fitness value then a predefined constant. Basically, CSCA has a training phase that will return a pool of antibodies used as a criterion to compute the affinity for unlabeled antigens in the detection stage. The classification decision is made as a result of the k best affinity matches and a majority vote for the class. Figure 1 describes the main stages of CSCA.

3. Proposed IDS model

The proposed IDS model is anomaly based and its architecture is based on two components: Local Collector (acquires logs from one host and classifies them as anomaly, suspicious or normal) and Central Collector (acquires all the suspicious data from multiple local collectors and decides if these are anomalies or not). A diagram of the proposed model is given in figure 1.

3.1 The local collector

This component is associated locally to the host and labels data as: anomaly (if MCAV > 0.85), suspicious (MCAV $\epsilon [0.65, 0.85]$) or normal (MCAV < 0.65), using DCA with antigen segmentation (Gu et al. 2009). If data is classified as an anomaly then the process generating the data is blocked. While, the suspicious data is transmitted to the central collector and if classified as an anomaly it blocks the originating process.

3.2 B. The central collector

All Local Collectors send their suspicious data to the Central Collector that will finally decide if it is an anomaly or not. This central unit will use the CSCA in order to determine if data is an anomaly or not. Furthermore, in order to compare it with other established classifiers we test the suspicious data with Support Vector Machines (SVM) and decision trees (C4.5).

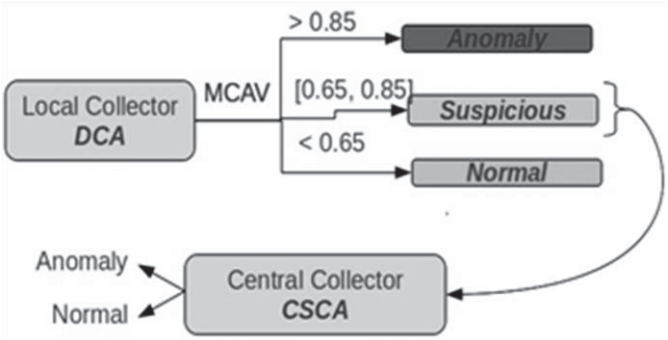


Figure 2: Proposed IDS model

4. Test results and analysis

We used the NSL-KDD dataset (Tavallaee et al. 2009) to test our model. Each record from the dataset has 41 features and is labeled as normal or attack. There are four categories of intrusion including: Denial of service (DoS), Remote- to-Local (R2L), User-to-Root (U2R) and Probing. From this dataset we randomly select 9,500 records. To categorise the signals we used Information Gain and obtained:

- PAMP: serror_rate, srv_serror_rate, same_srv_rate, dst_host serror and dst_host_rerror_rate;
- Danger: count and srv_count;

- Safe: logged_in, srv_different_host_rate and dst_host_count

Our experiments were conducted on a PC with Windows 8.1 operating system, CPU intel core i7 and 8 GB of RAM. We implemented in Java dDCA with antigen multiplication (AM) and antigen segmentation (AS). In order to evaluate them we used two experimental setups detailed in table 2. For the antigen segmentation we used three different segmentations in order to identify its influence on the sensitivity of the system's results. These approaches are denoted as: $AS_{i\cdot1}$ (100 antigens per segment), $AS_{i\cdot5}$ (500 antigens per segment) and $AS_{i\cdot9}$ (950 antigens per segment), where i is associated with the test case scenario (1 or 2). For the second scenario, with fewer cells, we include two extra segment sizes and designate them as: $AS_{2\cdot2}$ (20 antigens per segment) and $AS_{2\cdot0}$ (50 antigens per segment).

Table 2: Setup parameters for local collector

dDCA vers.	Nb. of cells	Antigen multiplication	Lifespan interval	Antigen segmentation
AM_1	100	10	$\in$ [100, 300]	
AM_2	20	5	$\in$ [20, 60]	
AS_1	100	10	$\in$ [100, 300]	$\in$ {100, 500, 950}
AS_2	20	5	$\in$ [20, 60]	$\in$ {100, 500, 950}
AS_2*	20	5	$\in$ [20, 60]	$\in$ {20, 50}

Results from table 3 show DCA achieves good results in terms of Attack Detection Rate (ADR), False Alarm Rate (FAR) and execution time (for dDCA with segmentation we took into account the processing time for one segment, because it is after this period that the algorithm offers a response). We take note that for our results we have considered only the anomalous and normal records, while the suspicious data will be evaluated by the Central Collector. Antigen multiplication is added in order to improve performances, as this implies a single antigen will be proliferated and presented for analysis to multiple DC cells. Hence, the antigen will not be classified based on the "judgement" of one single cell but, as an average of various cells that have different lifespan values. Moreover, segmentation can produce multiple sets of results, instead of one set as in the non-segmented case. Therefore, we tested the approaches 20 times and recorded the average value of ADR and FAR obtained after all segments had been processed.

Table 3: Test results for the local collector

dDCA vers.	ADR (%)	FAR (%)	Exec. Time (ms)	Suspicious records
AM_1	61 (+0.2/-0.3)	4.1 (+/-0.12)	$642X\ 10^3$	8,071(+/-10)
$AS_{1\cdot1}$	68 (+0.2/-0.34)	3 (+0.4/-0.2)	[200, 270]	4,250(+30/-9)
$AS_{1\cdot5}$	79 (+1.5/-0.5)	3.3 (+/-0.21)	[2600, 2705]	6,500(+20/-5)
$AS_{1\cdot9}$	72 (+1.2/-0.2)	4 (+0.11/-0.10)	[6120, 7050]	6,900(+/-20)
AM_2	71 (+0.9/-0.4)	5 (+/-0.25)	$16\ X\ 10^3$	2,250 (+/-30)
$AS_{2\cdot2}$ *	78(+1.2/-0.34)	8 (+0.9/-0.2)	[5, 8]	2,250(+40/-3)
$AS_{2\cdot0}$ *	80(+3.1/-0.41)	6 (+0.6/-0.4)	[8, 13]	2,220(+30/-5)
$AS_{2\cdot1}$	78(+1.2/-0.34)	5 (+0.6/-0.4)	[12, 16]	2,200(+34/-2)
$AS_{2\cdot5}$	76 (+1.5/-0.5)	5 (+/-0.31)	[15, 227]	2,070(+50/-5)
$AS_{2\cdot9}$	75 (+1.2/-0.1)	5.1 (+0.11/-0.1)	[225, 409]	2,010(+40/-3)

In the first scenario we setup dDCA with higher values for the lifespan interval and antigen multiplication factor, meaning an antigen might be tested more times than in the second case. AM_1 has poor results, as the execution time (10 mintues) is high and the number of suspicious data is almost 85% of the original dataset. Segmentation reduces the dataset to almost 45% in the case of $AS_{1\cdot1}$ (with segments of size 100) and improves ADR with 13% for $A_{1\cdot5}$, when compared to the multiplication approach (AM_1). As the number of antigens included in a segment increases, the ADR has a descending evolution, if the number of cells is higher than the number of antigens in the segment. This evolution is further captured by AM_1 where the segment size is equal to the number of records. If the number of DC cells is equal or lower than the segment size then the multiplication process is somewhat biased because it could be the same cell that will analyse a certain antigen; this explains the lower performances of $AS_{1\cdot1}$, when compared with $AS_{1\cdot5}$ that has a larger number of antigens. On the other hand, FAR has a slight descending trend when the segment size is raised, because as the number of antigens sent for analysis is lower DCA becomes more sensitive to false positives.

Regarding the second scenario, reducing the number of cells, antigen multiplication factor and lifespan thresholds, results in better performances with a lower number of suspicious data. AM_1 obtains good performances and with a low number of suspicious data, but the execution time is still significantly high (16 seconds). By adding segmentation to dDCA, the model obtains higher ADR (with a maximum of almost 12% when compared with AM_2) and faster response time (less then 10 ms in the best case). While varying the number of antigens in each segment, we acknowledge that segmentation does enhance ADR, when compared with AM_2. If the number of cells is lower then the number of antigens from the segment, then ADR increases, until the number of antigens exceeds the population of DC cells. In this last case larger number of antigens in segments will decrease ADR. As in the previous case, a smaller number of antigens considered for analysis will hinder FAR.

Table 4: Test results for the central collector

Algorithm	Suspicious records	ADR (%)	FAR (%)
Suspicious dataset - AM_1			
CSCA	8071	94.01	5.32
SVM	8071	98.22	1.52
C4.5	8071	99.11	0.080
Suspicious dataset - $AS_{1.9}$			
CSCA	6920	94.37	4.99
SVM	6920	91.60	3.67
C4.5	6920	99.19	0.07
Suspicious dataset - $AS_{1.1}$			
CSCA	4251	93.92	4.96
SVM	4251	90.40	6.90
C4.5	4251	99.04	0.077
Suspicious dataset - AM_2			
CSCA	2250	98.50	6.50
SVM	2250	81.32	62.03
C4.5	2250	99.64	0.155

In both experimental scenarios, segmentation will improve the ADR, while the number of false alarms slightly increases with a negligible value.

Finally the suspicious data, resulted from the local collector, is evaluated by the central component, with results shown in table 4. We considered only some of the suspicious datasets, such that each will have a different number of records. Moreover, the proportion of anomalies in the subsets is similar. To evaluate the central collector, we performed 10 fold cross validation and used the default version of the algorithms (SVM, J45 and CSCA) from weka version 3.6.10 (Hall et al. 2009). CSCA obtains good performances, outperforming SVM in most cases and obtaining performances close to the well-known C4.5. From the low scores obtained by SVM we can deduce that this classifier requires a larger dataset for learning, while CSCA and decision trees perform well even after less iteration for learning.

5. Conclusions and future work

In this paper we have exploited two AIS algorithm to construct an anomaly based IDS model with active response built on two main components: local and central collector. The first component will conduct a rapid analysis, near real-time, and label data as anomaly, suspicious or normal. The originating source is automatically blocked for traffics identified as anomalous, while suspicious data are sent to the central component for further analysis. Finally, the central component will receive the suspicious data and perform an off-line detection.

DCA has several limitations and requires adjustments for many parameters, which do not have a well-defined rule. Two version of DCA are implemented for the local component, including antigen multiplication and segmentation. Tests on the NSL-KDD dataset proved that multiplication in segments enhances ADR, but it slightly increases the number of false alarms. Therefore, DCA does not function well when a small number of antigens are presented. Furthermore, the segment size should exceed the number of DC cells, but if the size of the segment increases too much analysis will show poorer performances. Our experiments have shown that in both scenarios, having a smaller number of antigens will improve ADR, but will increase the number of false alarms. Our DCA version has two static thresholds for MCAV, because this is the boundary that could present an

undecided "area". Therefore, we send this data to the central component. Furthermore, we could have send all the data as being suspicious, but we considered this would burden network infrastructure; nonetheless, this might be a second approach.

For the Central Collector we used CSCA and compared it with other two well established machine learning classifiers: SVM and decision trees (C4.5). Results show that CSCA obtains similar performances as the other classifiers.

Future work will consider improving the DCA method by offering a feedback mechanism in order to improve its performance scores, by using a method such as Naive Bayes. Furthermore, the theoretical model can constitute the starting point to construct other tools for information security assurance including risk management, early based warning systems or visual analytics for cybersecurity, as all these types of solutions can benefit from our proposed theoretical model which succeeds to reduce the input dataset while amplifying the performance results.

References

Al-Hammadi, Y., Aickelin, U. & Greensmith, J. (2008), Dca for bot detection., in 'IEEE Congress on Evolutionary Computation', IEEE, pp. 1807–1816.

Brownlee, J. (2005), 'Clonal selection theory and clonalg (technical report no. 2–02)'.

Chelly, Z. & Elouedi, Z. (2010), Fdcm: A fuzzy dendritic cell method., in E. Hart, C. McEwan, J. Timmis & A. Hone, eds, 'ICARIS', Vol. 6209 of Lecture Notes in Computer Science, Springer, pp. 102–115.

de Castro, L. & Timmis, J. (2002), Artificial Immune Systems: A New Computational Intelligence Approach, Springer, Berlin, Heidelberg, New York.

Enache, A.-C. & Sgarciu, V. (2015), An improved bat algorithm driven by support vector machines for intrusion detection, in '8th International Conference on Computational Intelligence in Security for Information Systems', pp. 41–52.

Enache, A., Ionita, M. & Sgarciu, V. (2015), An immune intelligent approach for security assurance, in '2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)', pp. 1–5.

Europol (2015), 'Europol: The internet organised crime threat assessment (iocta) 2015' (https://www.europol.europa.eu/sites/2015/europol_iocta_web_2015.pdf)

Farmer, J. D., Packard, N. H. & Perelson, A. S. (1986), 'The immune system, adaptation, and machine learning', Phys. D 2(1-3), 187–204.

Forrest, S., Hofmeyr, S. A., Somayaji, A. & Longstaff, T. A. (1996), A sense of self for unix processes, in 'In Proceedings of the 1996 IEEE Symposium on Security and Privacy', IEEE Computer Society Press, pp. 120–128.

Forrest, S., Perelson, A. S., Allen, L. & Cherukuri, R. (1994b), Self- nonself discrimination in a computer, in 'In Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy', IEEE Computer Society Press, pp. 202–212.

Greensmith, J. . (2007), The Dendritic Cell Algorithm, PhD thesis, School of Computer Science, University of Nottingham.

Greensmith, J. & Aickelin, U. (2007), Dendritic cells for syn scan detection, in 'Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation', GECCO '07, ACM, New York, NY, USA, pp. 49–56.

Greensmith, J., Aickelin, U. & Twycross, J. (2006), Articulation and clarification of the dendritic cell algorithm, in 'Proceedings of the 5th International Conference on Artificial Immune Systems (ICARIS)', pp. 404–417.

Gu, F. (2011), Theoretical and Empirical extensions of the Dendritic Cell Algorithm, PhD thesis, Department of Computer Science, University College London.

Gu, F., Greensmith, J. & Aickelin, U. (2008), Further exploration of the dendritic cell algorithm: Antigen multiplier and time windows, in 'Proceedings of the 7th International Conference on Artificial Immune Systems (ICARIS)', IEEE, pp. 142–153.

Gu, F., Greensmith, J. & Aickelin, U. (2009), Integrating real-time analysis with the dendritic cell algorithm through segmentation, in 'Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation', GECCO '09, ACM, New York, NY, USA, pp. 1203–1210.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009), 'The weka data mining software: an update', SIGKDD Explor. Newsl. 11, 10–18.

Hofmeyr, S. A. & Forrest, S. (1999), Immunity by design: An artificial immune system, in 'Proceedings of the Genetic and Evo- lutionary Computation Conference', Vol. 2, Morgan Kaufmann, pp. 1289–1296.

Hunt, J. & Cooke, D. (1995), An adaptive, distributed learning system based on the immune system, in 'Proceedings of the IEEE International Conference on System, Man and Cyberspace', pp. 2494–2499.

Kim, J. & Bentley, P. J. (2001), Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator, in 'Proceeding of the Congress on Evolutionary Computation (CEC-2001)', pp. 1244–1252.

Kim, J. & Bentley, P. J. (2002), Towards an artificial immune system for network intrusion detection: An investigation of dynamic clonal selection, in 'Proceeding of the Congress on Evolutionary Computation (CEC-2002)', pp. 1015–1020.

Kim, J., Bentley, P. J., Aickelin, U., Greensmith, J., Tedesco, G. & Twycross, J. (2007), 'Immune system approaches to intrusion detection – a review', Natural Computing 6(4), 413–466.

Kim, J. W. (2002), Integrating Artificial Immune Algorithms for Intrusion Detection, PhD thesis, Department of Computer Science, University College London.

Kumari, K., Jain, A., Dongre, S. & Jain, A. (2012), 'Improving dendritic cell algorithm by dempster belief theory', International journal of Computer Engineering and Technology 2012 3(2), 415-423

Matzinger, P. (1994), 'Tolerance, danger and the extended family.', Annual Review of Immunology 12(1), 991–1045.

Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani, A. A. (2009), A detailed analysis of the KDD CUP 99 data set, in 'Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defense Applications', IEEE, pp. 1–6.

Continuous Supervision: A Novel Concept for Enhancing Data Leakage Prevention

Barbara Hauer
Johannes Kepler University, Linz, Austria
barbara_hauer@gmx.at

Abstract: The achievements of modern technology allow to access data and information independent of time and place. For business applications ever more sophisticated mobile devices, such as notebooks and smartphones, offer the opportunity to work outside the office in almost the same manner as on the business premises of an organization. But this advantages of accessing sensitive data without restriction of the location involves severe information security risks. Hence, information exposure, which is in the focus of data leakage prevention (DLP) and information leakage prevention (ILP), is a decisive key factor. Current state of the art enterprise content-aware DLP solutions offer different approaches to monitor and to protect confidential data at client endpoints. However, these DLP solutions are restricted to prevent technologically based data breaches and they are subject to several issues and limitations. One of these issues is the reliance on a single successful authentication of an authorized user. Therefore, this work introduces a novel concept for enhancing DLP on client endpoints by establishing continuous supervision for verifying the identity and the attendance of the user as well as for enhancing the privacy by preventing shoulder surfing. This concept is bases on mobile sensor technology and biometric methods which are used by state of the art mobile devices for security features as well as for smart features. In this context, a common DLP approach is combined with continuous supervision by utilizing the front camera of the mobile device. In order to prove the concept, this work presents a web technology based prototype which allows a wide range of application. In general, the implementation targets a high used user acceptance which is influenced by the performance efficiency as well as the detection reliability. Therefore, these benchmark figures are evaluated for various JavaScript libraries processed on mobile devices with different hardware specifications. The presented and discussed results indicate the technical feasibility of this concept. However, there are certain requirements with respect to the hardware specification of the mobile devices.

Keywords: information and data security, information and data leakage prevention, information exposure, continuous supervision, biometrics

1. Introduction

These days various challenges within the scope of data leakage prevention (DLP) and information leakage prevention (ILP) refer to mobile devices which have evolved to an inherent part of our daily life. A recent study launched by the German “Gesellschaft für Unterhaltungselektronik” (GFU) has shown that a significant number of employees access their business e-mails during leisure time (GFU Consumer & Home Electronics GmbH, 2015). Expressed in terms of numbers, 58 % of the Swiss employees, 56 % of the Italian employees, 54 % of the Austrian employees, 45 % of the Spanish employees, and 42 % of the German employees access their business e-mails outside the office. Furthermore, more sophisticated mobile devices, such as notebooks and smartphones, offer the opportunity to work outside the office in almost the same manner as on the business premises of an organization. But this advantage of accessing business data without restriction of the location involves severe information security (IS) risks. In this context, information exposure (CWE, 2015) is one of the most crucial points. Even if confidential data is transferred over an encrypted network connection, displayed in a secured environment such as a sandbox, and a DLP agent is utilized, this data can be exposed to an unauthorized person in case of misusing privileges by exploiting an unlocked system or by shoulder surfing. This is due to the fact that state of the art DLP solutions rely on a single successful authentication of an authorized user. However, mobile devices implement a variety of sensors which can be utilized by biometric techniques. Currently, most of these techniques are used to provide smart features while identity verification plays a minor role. For example, most mobile devices embed a front camera which can be used for video conferences. Furthermore, the camera allows to control games and movie players by eye tracking and recognizing face movements. Applications for eye control are already implemented in the android operating system and the Samsung Galaxy S4 was the first smartphone to provide the Samsung Smart Pause and Samsung Smart Scroll features (Samsung Electronics Co. Ltd., 2015). But the sensor technologies in mobile devices and their corresponding techniques provide new opportunities for DLP as well. In order to reduce the risk of exposing confidential data, DLP solutions can be combined with mobile sensor technologies and biometric techniques to continuously monitor the identity and the attendance of an authenticated and authorized user during data access and usage.

2. Continuous supervision using biometric identification methods

Biometrics, used for the identity verification of a person, are classified into physiological and behavioral biometrics depending on their characteristics. Physiological biometrics rely on the unique physical attributes of a person such as the fingerprint, face, iris, voice, hand geometry, ear, odor, retina, or vein (Jain, Ross and Prabhakar, 2004). The behavioral biometrics of a person, on the other hand, are based on the behavior of performing particular tasks. Relevant examples are the signature, keystroke, gait, gesture, or speech (Sujithra and Padmavathi, 2012). Due to the given limitations of hardware and software, certain biometrics are more suitable to be embedded in a mobile device than others. Face recognition, for example, is commonly used to verify and to identify a person. A variety of these approaches can be found in literature such as "Real-time face verification for mobile platforms" (Jung et al, 2008), "A new method for combined face detection and identification using interest point descriptors" (Stein and Fink, 2011), "Fast face recognition technique for small and portable devices" (Zaeri, Mokhtarian, and Cherri, 2006), or "A fast face recognition system on mobile phone" (Chen, Shen and Sun, 2012). Furthermore, there are approaches for eye tracking, e.g. "Reducing shoulder-surfing by using gaze-based password entry" (Kumar et al, 2007), "A geometric approach to remote eye tracking" (Villanueva, 2009) or "Robust real-time multi-user pupil detection and tracking under various illumination and large-scale head motion" (Yan, Wang and Zhang, 2011). Even if those techniques do not address the problem of continuous supervision of a person during confidential data access and usage, advanced methods provide continuous face tracking and automatic screen locking or unlocking for users. Sensible Vision, Inc., for example, offers a secure access system for electronic devices which applies continuous facial image tracking whenever the user physically moves into the camera's field of view (Azar and Brostoff, 2013).

In the context of DLP applications, biometrics can be used for continuous supervision of a person during data access and usage. The primary objective of DLP is to protect or at least to monitor confidential data. Various present limitations of DLP solutions can be resolved by a continuous supervision of the identity and the attendance. Biometric techniques seem to be advantageous due to the existence of mobile sensor technologies such as front cameras, microphones, touch screens, motion sensors, infrared sensors, and gravity sensors. In general, the quality of these mobile sensors is continuously increasing while the devices are miniaturized and the costs are reduced. Furthermore, mobile sensors are implemented in most mobile devices. Even today, several companies try to exploit mobile sensor technologies to provide advanced security features. Apple, Inc., for example, embeds a fingerprint sensor as well as a near field communication (NFC) technology in the iPhone 6 (European Experts in Electronic Transaction Systems (ESTEL), 2014). These features are utilized to secure Apple Pay which is a mobile payment and digital wallet service. However, mobile sensor technologies are exposed to threats associated with networks, applications, and malware. Therefore, it has to be assured that the security features function correctly even if the mobile sensors are bypassed or not working properly.

Several other aspects of surveillance and monitoring, such as the user acceptance and cooperation, have to be considered as well. Various research results, e.g. Grant and Higgins (1991), Wang (2010), and Al-Omari, El-Gayar and Deokar (2012), indicate that employees are willing to accept computerized monitoring and continuous surveillance and to comply with IS policies and measures under certain conditions. These conditions include factors such as the design and implementation of the surveillance system and the users' experience with these systems. Therefore, solutions providing DLP on mobile devices have to meet specific quality characteristics to ensure acceptance and cooperation.

3. DLP prototype for web applications including continuous supervision

Current DLP solutions are limited to a single successful authentication of an authorized user which results in a remaining risk of information exposure during data access and usage. Therefore, this work proposes to use mobile sensor technologies and biometric techniques for continuous supervision within the scope of DLP. The basic structure of this concept is shown in Figure 1. The draft combines a common DLP approach with continuous supervision of the identity and the attendance by utilizing the front camera of the mobile device. In the first steps, the camera is used for biometric identity verification to confirm that a person impersonates the successfully authenticated and authorized user. Subsequently, the camera is required to continuously supervise the identity and the attendance of the person in front of the mobile device while confidential data is displayed and used. Furthermore, the privacy is enhanced by detecting and preventing shoulder surfing. This requires an alignment of the display view and the camera's field of view by utilizing privacy filters and, potentially, wide-angle lens. Using this concept, security threats caused by unlocked systems can be avoided, too.

In general, the DLP prototype implements the entire model illustrated in Figure 1. Due to the fact that the prototype targets platform independent web applications, the client side implementation is based on Hypertext Markup Language (HTML) 5 and JavaScript (JS).

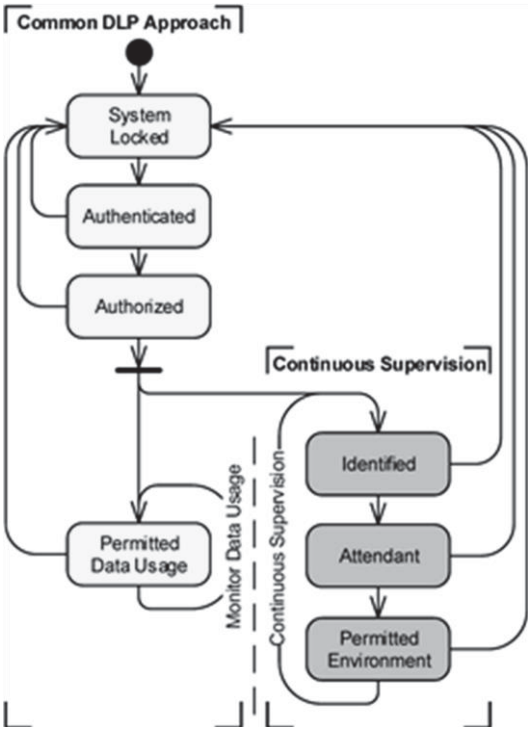


Figure 1: Basic structure of the DLP prototype for web applications including continuous supervision

In order to implement the common DLP approach, the client DLP application uses the front camera to record images of the person in front of the mobile device. These images as well as the entered login credentials are transmitted to a server. The server side implementation of the common DLP approach is based on the opencv library (Itseez, 2015) which is embedded in a dynamic link library (DLL) for fast identity verification. In case of a successfully authenticated and authorized user, the server provides the confidential data to the client DLP application. From now on, this application is responsible for the data usage monitoring and the continuous supervision. The implementation of the continuous supervision makes use of JS computer vision libraries such as ccv (Liu, 2015) and tracking.js (Lundgren et al, 2015). These libraries provide face and eye detection functionalities which are required to monitor the environment. If, for example, the face of the user is leaving the cameras field of view, the client DLP application locks the access to confidential data. Detecting multiple faces is an indication for shoulder surfing and leads to locking, too. After detecting a permitted environment again the client DLP application repeats the user identity verification by transmitting user images to the server. If the person is successfully identified as the authorized user the access to confidential data is unlocked.

Literature provides a variety of research result, e.g. Chan and Kittler (2010), Yacoob and Davis (2006), and Moura, Gomes and de Carvalho (2013), about the reliability and detection rates of algorithms for facial identity verification such as the one used for the server side implementation of the common DLP approach. Therefore, the benchmarking results presented in this work focus on the continuous supervision which is implemented in the client DLP application. These results are illustrated in the following section.

4. Benchmarking

The proposed DLP prototype focuses on two software product quality characteristics of the ISO/IEC 25010:2011 standard (ISO/IEC JTC 1/SC 7, 2011): the reliability, which is mainly represented by the detection reliability, as well as the performance efficiency which is an important criterion for mobile devices. This is due to the fact that these devices are commonly affected by limitations of the central processing unit (CPU) performance, the main memory size, the network bandwidth, and the energy source.

4.1 Test data and test environment

In order to provide comprehensible and repeatable tests, the test data consists of openly accessible components which are adapted to the desired use. The Database of Faces (AT&T Laboratories Cambridge, 2015), formerly known as The Olivetti & Oracle Research Laboratory (ORL) Database of Faces, contains ten different images of each of forty individuals. These frontal face images provide various facial expressions and facial details as well as different lighting. The images are available in a PGM file format with a size of 92x112 pixels and 256 levels of grayscale. In order to provide the required test data for the prototype, each image is converted to the PNG file format and cropped to its containing head. Based on these images the test data is generated by combining the facial images and background images which contain circles and rectangles in various shapes and sizes. This background provides a harsh test environment since it is more difficult to detect a face or an eye within similar geometric shapes. Additionally, the variation of the image and face size enables scenarios closer to the reality. For each combination of a certain image and face size 100 different images are analyzed to average the benchmarking results.

In general, the benchmarking test environment consists of a multi-tier architecture which includes an Apache Tomcat v7.0.50 (64-bit) web server. The client DLP application is analyzed on three different environments with the following specifications.

- Laptop client: HP EliteBook 8470p embedding an Intel Core i5-3320M 2.60 GHz CPU and 8 GB DDR3 RAM running Microsoft Windows 8 Pro (64-bit) and a Google Chrome web browser v.44.0.
- Smartphone client: HTC Desire HD embedding a 1 GHz Scorpion CPU and 768 MB RAM running Android v.4.4.2 API 19 and a Google Chrome web browser v.43.0.
- Smartphone client: LG G2 embedding a Quad-core 2.26 GHz Krait 400 CPU and 2 GB RAM running Android v.4.4.2 API 19 and a Google Chrome web browser v.43.0.

The selected clients demonstrate the impact of the hardware specification on the performance efficiency, especially the execution time. Additionally, interfering software factors, such as web browser cache, antivirus software, firewall software, and other client software not needed for benchmarking, are disabled in order to avoid interferences of the measurement results.

The JS heap size as well as the CPU execution time of the algorithms are analyzed using Google Chrome DevTools (Google, Inc., 2015). This tool set also allows to execute and analyze the test cases on the smartphone clients by utilizing remote debugging via Google Chrome.

4.2 Detection reliability

The detection reliability expresses the ability to detect faces or eyes which is a required functionality to implement continuous supervision. The detection itself does not imply the identification or verification of an individual person. Therefore, the positive predictive value (PPV), which represents the probability that positive predictions are truly correct, is more important for continuous supervision than the prediction accuracy (ACC). In general, the true positive rate (TPR), the false positive rate (FPR), and the true negative rate (TNR) have to be high, and the false negative rate (FNR) has to be low. This is due to the fact that confidential data has to be protected in case of uncertainty. Furthermore, the TPR has to be prioritized over the PPV when it comes to continuous supervision because it is even more important to detect present faces or eyes than to detect faces or eyes without faults. On the other hand, the user acceptance is affected negatively if the TNR is too low. In contrast to the continuous supervision, the algorithms for identifying or verifying a person have to prioritize both the TPR and the TNR in order to not provide data access to an unauthorized individual.

Figure 3, Figure 4, and Figure 5 depict the FNR, the TPR, and the PPV of the JS libraries as a function of the IWH and the FH given in pixel. It can be seen that the high FNR of the tracking.js library used for eye detection correlates to the insufficient ACC for this test data. Figure 4 and Figure 5 illustrate similar relations for the TPR and the PPV. The TPR results have a remarkable resemblance to the ACC results because the test data is only based on so called positive examples. Negative examples, which do not contain faces or eyes, are not within the focus of this benchmarking. Accordingly, the PPV results shown in Figure 5 differ significantly from the ACC and TPR results. For the test data ”IWH 160x200; FH 200”, ”IWH 480x320; FH 150”, and ”IWH 1280x720; FH 150” the tracking.js library used for eye detection reaches a PPV nearly as high as the PPV of the JS libraries used for face detection. In summary, the ccv library shows nearly constant ACC, TPR, FNR, and PPV results which are almost

not affected by the IWH and the FH. An exception are too small faces which have a negative effect on the ACC, TPR, and FNR. The benchmarking results of the tracking.js library, on the other hand, are notably dependent on the ratio between the IWH and the FH as well as on the FH itself. In general, the TPR, FNR, and PPV results interrelate with the ACC results of the test data. Accordingly, a low detection ACC negatively affects other rates. Nevertheless, the ACC itself is not enough to evaluate a library and the underlying algorithm because a high ACC does not necessarily imply a high TPR and a high PPV.

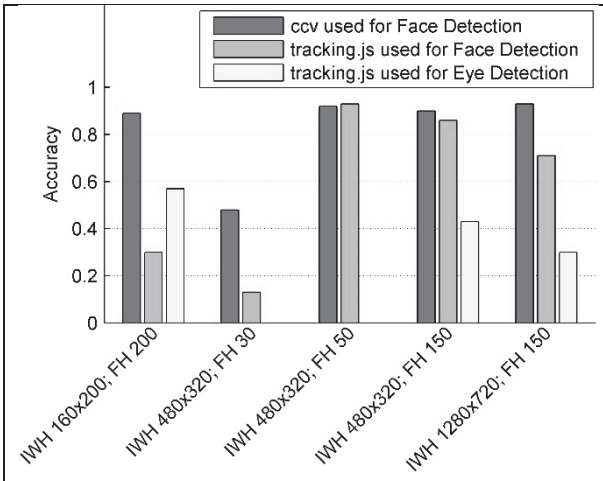


Figure 2: ACC of the JS libraries dependent on the IWH and FH

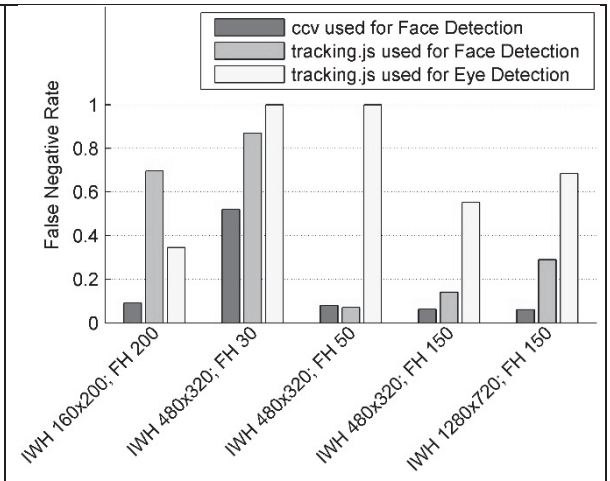


Figure 3: FNR of the JS libraries dependent on the IWH and FH

Figure 2 illustrates the detection ACC of the ccv and tracking.js library as a function of the image width and height (IWH) and the face height (FH) given in pixel. The results indicate that the ACC decreases if the face is too small or too large in relation to the IWH. The impact of this ratio can be examined by comparing, for example, the ACC of the test data “IWH 480x320; FH 30” and “IWH 480x320; FH 50”. Figure 2 also demonstrates that the tracking.js library requires a minimum FH of 150 pixels to correctly detect eyes at all.

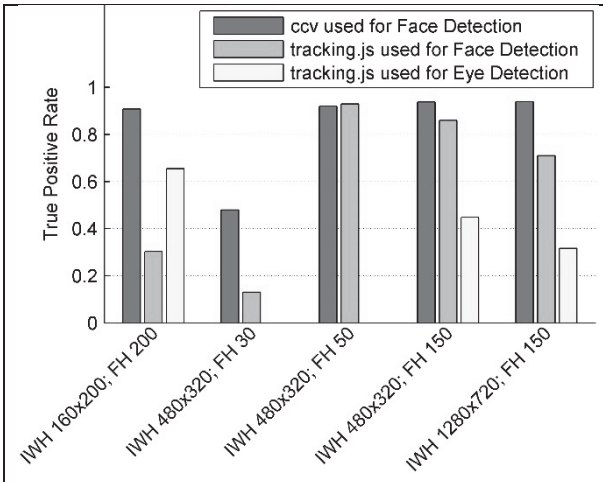


Figure 4: TPR of the JS libraries dependent on the IWH and FH

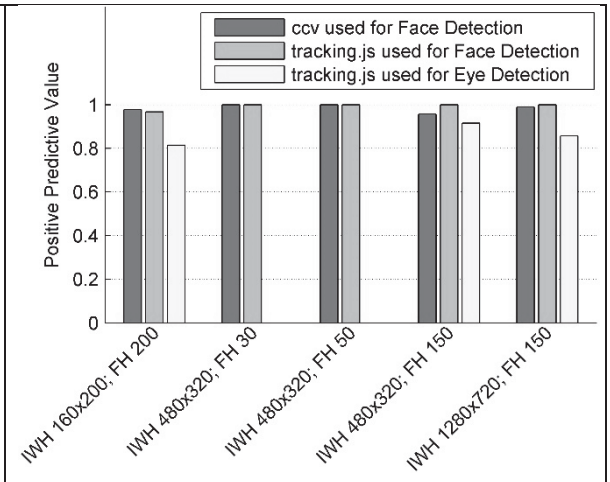


Figure 5: PPV of the JS libraries dependent on the IWH and FH

4.3 Performance efficiency

According to ISO/IEC JTC 1/SC 7 (2011), the performance efficiency involves the resource utilization, the time behavior, and the capacity. The benchmarking results presented in this work focus on the CPU execution times and the JS heap sizes for executing the JS libraries ccv and tracking.js in the Google Chrome web browser. In this context, the CPU execution time is one of the most critical factors since it directly influences the frame rate at which continuous supervision can be performed. Field tests have shown that a frame rate of approx. two images per second is sufficient to correctly verify the identity of the user and to detect shoulder surfing. In order to minimize interferences on the measurement results, e.g. caused by caching or heap management, a custom built JS function successively analyses a cluster of ten images selected from the test data. Therefore, the presented

measurement results do not refer to the processing of a single image. They represent the processing of ten images. This is an important fact, especially for the consideration of execution times.

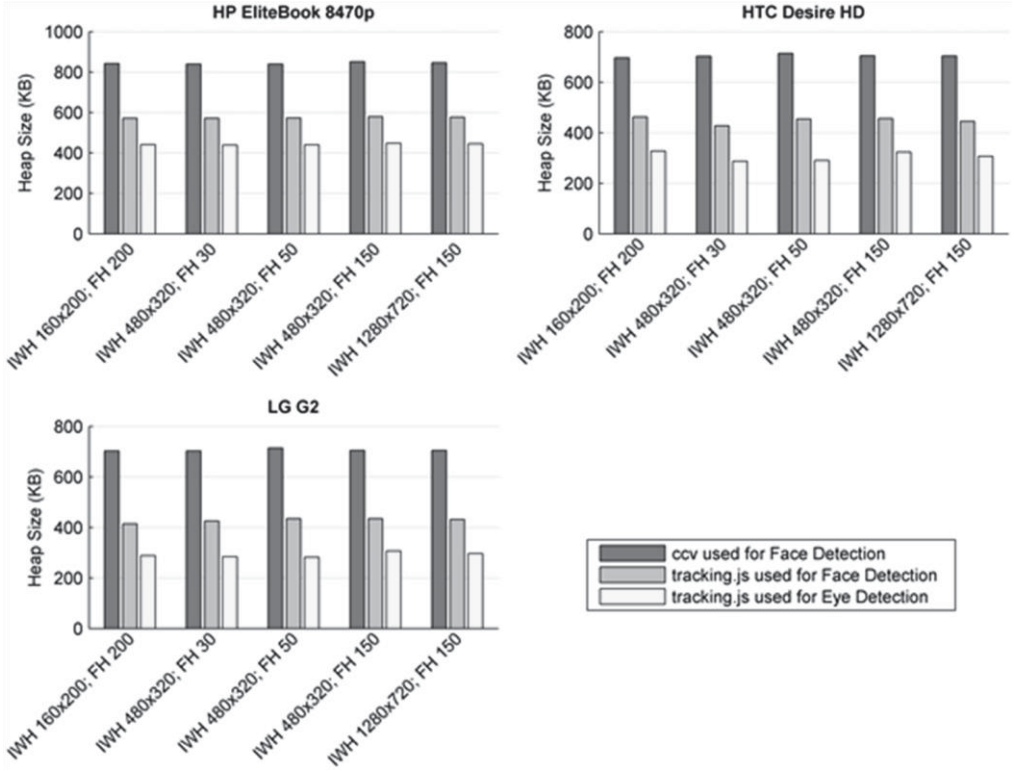


Figure 6: Average retained heap size of the JS libraries dependent on the IWH and FH

The measurement results in Figure 6 present the average retained JS heap size of the parent window which contains the custom built JS function for calling the corresponding function of the ccv and tracking.js library. The diagrams depict the average retained heap size of the JS libraries as a function of the IWH and the FH given in pixel. It can be seen that the heap sizes of the ccv library are extensively higher than the heap sizes of the tracking.js library. Expressed in terms of numbers, the heap size of the ccv library is approx. 30 % to 40 % higher than the heap size of the tracking.js library used for face detection and approx. 50 % to 60 % higher than the heap size of the tracking.js library used for eye detection. In addition, the measurement results indicate that the heap sizes are almost independent of the IWH and the FH. Furthermore, it can be seen that the heap sizes of the JS libraries are approx. 20 % to 40 % lower on the smartphone clients than the heap sizes on the laptop client. This is due to the fact that the smartphone clients provide less main memory and therefore, the JS engine has to implement different heap management strategies.

The CPU execution time is another important benchmark figure. Therefore, the JS libraries are evaluated by comparing the aggregated total time which includes all calls to the corresponding functions of the ccv and tracking.js library. The diagrams in Figure 7 demonstrate the average execution time of the JS libraries as a function of the IWH and the FH given in pixel. As stated above, the execution times refer to the processing of ten images. It can be seen that the average execution times mainly depend on the IWH whereas the influence of the FH is very low. For the laptop client the execution times of the tracking.js library used for face detection are approx. 30 % to 50 % lower than the execution times of the ccv library. In case of increasing the IWH from 480x320 pixels to 1280x720 pixels, the execution times of both JS libraries are roughly doubled. These results demonstrate that the JS libraries allow a sufficient frame rate of approx. three to six images per second for an IWH of 480x320 pixels on the laptop client. However, the frame rate drops below two images per second for an IWH of 1280x720 pixels.

In order to demonstrate the impact of the hardware specification on the execution time, the JS libraries are analysed using two different smartphone clients, too. The diagrams in Figure 7 indicate that the execution times increase tremendously on an outdated smartphone client which is represented by the HTC Desire HD. Compared to the laptop client, the average execution times of the tracking.js libraries are approx. six to twelve times longer. For the ccv library this factor is even higher, reaching a value of approx. 19 to 32. This results in frame rates

below two images per second for all IWH and JS libraries and hence, continuous supervision is restricted on this smartphone client. The impact on the execution time is less significant on smartphone clients embedding advanced hardware such as the LG G2. Compared to the laptop client, the average execution times of the tracking.js libraries are approx. three to four times longer and the average execution times of the ccv library are approx. four to eight times longer. Subsequently, the tracking.js libraries allow a sufficient minimum frame rate of two images per second for an IWH of 480x320 pixels and below. The ccv library, on the other hand, is not suitable for continuous supervision on this smartphone client.

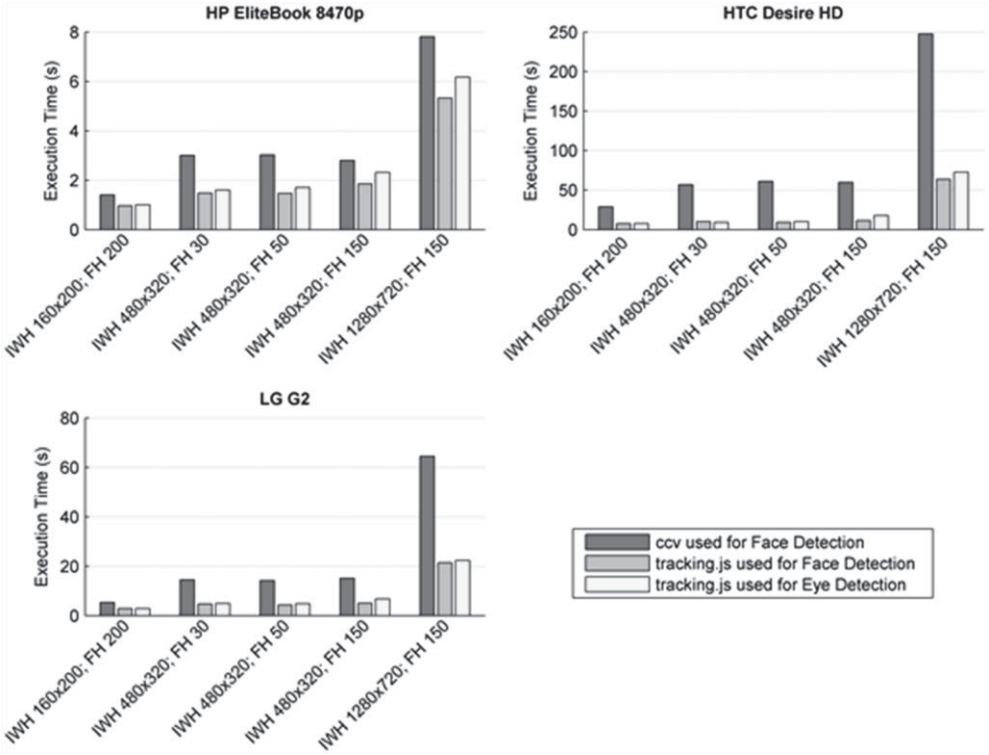


Figure 7: Average CPU execution time of the JS libraries dependent on the IWH and FH

In general, the measurement results demonstrate that continuous supervision using JS libraries heavily depends on the hardware specification of the client device. Hence, it can be restricted on outdated smartphone clients.

5. Conclusion

Due to the increasing impact of mobile devices in our daily life, DLP and ILP have evolved to an important part in this domain. However, current state of the art enterprise content-aware DLP solutions are subject to several limitations, one of them being the reliance on a single successful authentication of an authorized user. Hence, this work presents a novel concept for enhancing DLP on client endpoints by establishing continuous supervision. The approach allows to verify the identity and attendance during data usage and enhances the privacy by preventing shoulder surfing. In order to prove the concept, a web technology based prototype is presented and analysed. The benchmarking results of the JS libraries, which are utilized for the client DLP application, illustrate a sufficient detection reliability assuming that the face has an appropriate size. Considering the performance efficiency, the execution time is heavily depending on the hardware specification of the client device. This can lead to restrictions in case of applying the JS libraries for continuous supervision on outdated smartphone clients.

References

Al-Omari, A., El-Gayar, O. and A. Deokar, A. (2012) “Security policy compliance: User acceptance perspective” in Proceedings of the 45th Hawaii International Conference on System Science (HICSS), Maui, HI, USA, January 2012, pp. 3317–3326.

AT&T Laboratories Cambridge (2015) “The Database of Faces”, [Online], http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Azar, C. and Brostoff, G. (2013) “System and method for providing secure access to an electronic device using continuous facial biometrics”, US Patent No. US8370639 B2, Filed June 16th., 2005, Issued February 5th., 2013.

Chan, C.-H. and Kittler, J. (2010) “Sparse representation of (multiscale) histograms for face recognition robust to registration and illumination problems” in Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), Hong Kong, September 2010, pp. 2441–2444.

Chen, B., Shen, J. and Sun, H. (2012) “A fast face recognition system on mobile phone” in Proceedings of the 2012 International Conference on Systems and Informatics (ICSAI), Yantai, China, May 2012, pp. 1783–1786.

CWE (2015) “CWE-200: Information Exposure”, [Online], http://cwe.mitre.org/data/definitions/200.html

European Experts in Electronic Transaction Systems (ESTEL) (2014) “Apple iPhone 6, Apple Pay, What else?”, October 2014.

GFU Consumer & Home Electronics GmbH (2015), “Immer weniger Trennung von Freizeit und Arbeit”, August 2015, [Online], http://www.gfu.de/presseraum/uebersicht/immerweniger-trennung-von-freizeit-und-arbeit/

Google, Inc. (2015) “Chrome DevTools”, [Online], https://developers.google.com/web/tools/chrome-devtools

Grant, R. and Higgins, C. (1991) “Computerized performance monitors: Factors affecting acceptance”, IEEE Transactions on Engineering Management, vol. 38, no. 4, November 1991, pp. 306–315.

Hauer, B. (2014) “Data leakage prevention - A position to state-of-the-art capabilities and remaining risk” in Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS), vol. 2, Lisbon, Portugal, April 2014, pp. 361–367.

ISO/IEC JTC 1/SC 7 (2011) “ISO/IEC 27001:2011: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models”, March 2011.

Itseez, (2015) “Open Source Computer Vision Library (OpenCV)”, [Online], http://http://opencv.org

Jain, A., Ross, A. and Prabhakar, S. (2004) “An introduction to biometric recognition”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, January 2004, pp. 4–20.

Jung, S.-U., Chung, Y.-S., Yoo, J.-H. and Moon, K.-Y. (2008) “Real-time face verification for mobile platforms” in Advances in Visual Computing, ser. Lecture Notes in Computer Science, vol. 5359, Springer Berlin Heidelberg, 2008, pp. 823–832.

Kumar, M., Garfinkel, T., Boneh, D. and Winograd, T. (2007) “Reducing shoulder-surfing by using gaze-based password entry” in Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS), Pittsburgh, PA, USA, July 2007, pp. 13–19.

L. Liu, (2015) “ccv - A Modern Computer Vision Library”, [Online], http://libccv.org/

Lundgren, E., Rocha, T., Rocha, Z., Carvalho, P. and Bello, M. (2015) “tracking.js - A modern approach for Computer Vision on the web”, [Online], http://trackingjs.com

Moura, E., Gomes, H. and de Carvalho, J. (2013) “An improved face verification approach based on speedup robust features and pairwise matching” in Proceedings of the 26th Conference on Graphics, Patterns and Images (SIBGRAPI), Arequipa, Peru, August 2013, pp. 362–369.

Samsung Electronics Co. Ltd. (2015) “Samsung GALAXY S4”, [Online], http://www.samsung.com/global/microsite/galaxys4

Stein, S. and Fink, G. (2011) “A new method for combined face detection and identification using interest point descriptors” in Proceedings of the 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG), Santa Barbara, CA, USA, March 2011, pp. 519–524.

Sujithra, M. and Padmavathi, G. (2012) “Next generation biometric security system: An approach for mobile device security” in Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology (CCSEIT), Coimbatore UNK, India, October 2012, pp. 377–381.

Villanueva, A., Daunys, G., Hansen, D., Bohme, M., Cabeza, R., Meyer, A. and Barth, E. (2009) “A geometric approach to remote eye tracking”, Universal Access in the Information Society, vol. 8, no. 4, November 2009, pp. 241–257.

Wang, P. A. (2010) “Information security knowledge and behavior: An adapted model of technology acceptance” in Proceedings of the 2nd International Conference on Education Technology and Computer (ICETC), vol. 2, Shanghai, China, June 2010, pp. 364–367.

Yacoob, Y. and Davis, L. (2006) “Detection and analysis of hair”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, July 2006, pp. 1164–1169.

Yan, C., Wang, Y., and Zhang, Z. (2011) “Robust real-time multi-user pupil detection and tracking under various illumination and large-scale head motion”, Computer Vision and Image Understanding, vol. 115, no. 8, August 2011, pp. 1223–1238.

Zaeri, N., Mokhtarian, F. and Cherri, A. (2006) “Fast face recognition technique for small and portable devices” in Proceedings of the 2006 IEEE International Workshop on Imaging Systems and Techniques (IST), Minori, Italy, April 2006, pp. 114–118.

E-CMIRC: Towards a Model for the Integration of Services Between SOCs and CSIRTs

Pierre Jacobs[1, 2], Sebastiaan von Solms[2] and Marthie Grobler[1, 2]
[1]Council for Scientific and Industrial Research, South Africa
[2]University of Johannesburg, South Africa
pjacobs@csir.co.za
basievs@uj.ac.za
mgrobler1@csir.co.za

Abstract: Security Operation Centres (SOCs) and Computer Security Incident Response Teams (CSIRTs) or Computer Emergency Response Teams (CERTs) can play a pivotal role in the monitoring of, and response to threats, attacks and vulnerabilities in organisations, including governments. While the focus of a SOC is on the monitoring of technical security controls and critical assets, and the response to attacks and threats, CSIRTs' main focus is on response and incident management. One postulation is that a CSIRT or CERT is a highly specialised sub-capability of a SOC, whereas another postulation could be that a SOC serves as an input mechanism into CSIRTs and CERTs. In this paper, the differences between SOCs, CERTs and CSIRTs are established, and synergies between them are defined. This leads to an integrated services model for the establishment of an initial SOC and CSIRT capability in developing countries. Developing countries have unique challenges facing them where it concerns cybersecurity. Aspects such as Information Communication and Technology (ICT) infrastructure are often a challenge, and so is funding for ICT as well as skills. Political instability could also have an influence on the cybersecurity posture of developing countries by leaving developing nations open to malicious state-sponsored attacks. This SOC and CSIRT capability is made viable and possible through the savings in cost and resources by identifying overlapping services, as well as the application of the proposed model. This emergent SOC and CSIRT combined capability is called the Embryonic Cyberdefense Monitoring and Incident Response Center (E-CMIRC). The purpose of this paper is to identify a high-level integrated services model for the E-CMIRC in order to reduce cost and resources which serves as a barrier to entry in developing countries. A scalable operational framework is identified, and for the management of the effectiveness and efficiency, and also to ensure that all aspects of service delivery are considered, the Information Technology Information Library (ITIL) is proposed.

Keywords: SOCs, CSIRTs, developing countries, security, service integration

1. Introduction

This paper aims to explore the differences and similarities between Security Operation Centres (SOCs) and Computer Security Incident Response Teams (CSIRTs). Services offered by SOCs and CSIRTs are in essence the same, with the biggest differentiators being scope, reach and technical monitoring capability. A recent book by Zimmerman (2014) makes little distinction between the services provided by SOCs and CSIRTs, and uses the terms interchangeably.

In this paper, services applicable to both SOCs and CSIRTs are identified. A model is proposed for the integration of these identified services for deployment in developing countries. This will be achieved by defining the functions and services of SOCs and CSIRTs. SOC and CSIRT services will then be grouped by scope or reach, and technology. Section 2 provides the background to the description and purposes of SOCs and CSIRTs. The SOC functions are extrapolated from Zimmerman (2014) and IBM (2016), and the CSIRT functions from IETF (Danyliw, 2004) and SANS (Campbell, 2003). Section 3 presents an extrapolation of SOC services as described by Zimmerman (2014), industry in South Africa (Jacobs, 2015), CSIRT services as described by The European Network and Information Security Agency (ENISA) (2015a) and the Carnegie-Mellon University Software Engineering Institute (SEI) (SEI - Carnegie Mellon, 2002). This culminates in a table presenting services both common and unique to SOCs and CSIRTs.

In Section 4 the SOC and CSIRT models, scope and reach is compared, as well as technology common and unique to both. The Embryonic Cyberdefence Monitoring and Incident Response Center (E-CMIRC) service integration model is proposed in Section 5 while Section 6 proposes future work. The paper is concluded in Section 7.

2. Cyber challenges faced by developing countries

Most developing countries are adopting Information and Communications Technology (ICT) as an economic enabler, and to form part of the larger national internet community. Developing countries are however exposed

to unique challenges where it comes to ICT security. Some of these challenges are (Tagert, 2010) (Ghernaouti-Hélie et al, 2007):

- Complex government structures.
- Different cultures and levels of education in terms of ICT usage.
- Fiscal challenges.
- Resources and training.

Securing the cyber assets of developing countries is important to allow them to reap the benefits of ICT and the internet based economy. It also protects the developing state form nation sponsored attacks and cyber espionage (NATO Cooperative Cyber Defence Centre of Excellence, 2012). As such, the establishment of a SOC and/or CSIRT is of importance to assist in this regard.

3. CSIRT and SOC functions

There is a difference between the terms CSIRT and CERT or Computer Emergency Response Team Coordination Center (CERT/CC). In the development of the E-CMIRC model, these terms are used interchangeably. In general, the CSIRT abbreviation is mostly used in Europe for the protected CERT, or CERT/CC name. The name "CERT" and "CERT/CC" are however registered and owned by Carnegie Mellon University, where the first CSIRT was instituted (SEI at Carnegie Mellon University, 2015). The CERT/CC forms a sub-component of the larger CERT division. Many CSIRTs have been allowed by Carnegie Mellon University to use the name CERT or

CERT/CC in their names, but these are independent of the University (SEI at Carnegie Mellon University, 2015).

CERTs and CSIRTs both are made up of teams responding to cybersecurity incidents. The terms CSIRT is used as a generic description of an incident response team (Killcrece *et al*, 2003), while CERT is a trademarked name which is controlled by CERT/CC (Tagert, 2010).

3.1 SOC functions

Literature review by the authors on publicly available sources did not turn up comparative information on the differences of functions offered by SOCs and CSIRTs. However, during the European Chief Information Security Officer (CISO) Information Security Workshop in 2013, it was concluded that a SOC's responsibility spans the daily operations of security events (Bancroft *et al*, 2013). In support of this, a CSIRT's responsibility pertains to emergency response, investigation and *"the development of IT security architectural and engineering solutions"* (Experts Exchange, 2013). Accordingly, at a high level, the major differences between SOCs and CSIRTs can be categorised according to:

- Functions and services (refer to Section 4).
- *People and skills.*
- *Technology necessary to support those functions and services.*
- Scope or reach (refer to Section 5).

Zimmerman (2014) defines a SOC as "*a team primarily composed of security analysts organized to detect, analyze, respond to, report on, and prevent cybersecurity incidents"*, while Kelley and Moritz (Kelley *et al*, 2006) define one of its functions as to *"…monitor(s) and manage(s) all aspects of enterprise security in real-time from a single, centralized location".* Carnegie Mellon University defines a CSIRT as *"a service organisation that is responsible for receiving, reviewing, and responding to computer security incident reports and activity. Their services are usually performed for a defined constituency that could be a parent entity such as a corporate, governmental, or educational organisation; a region or country; a research network; or a paid client"* (SEI at Carnegie Mellon University, 2015). ENISA defines a CSIRT as "*a team of IT security experts whose main business is to respond to computer security incidents. It provides the necessary services to handle them and support their constituents to recover from breaches"* (ENISA, 2015b).

In determining the services applicable to SOCs and CSIRTs and possible duplication of those services, SOC and CSIRT functions and services are defined, and then grouped by scope (or reach) and technology. This approach will assist in determining where there is an overlap of services between the SOCs and CSIRTs. It will also contribute to the creation of the proposed E-CMIRC model in that it will serve as the first order integrated service

identification applicable to SOCs and CSIRTs. The purpose or activity for which SOCs exist, as taken from Zimmerman (2014) and IBM (2016) is shown in Table 1. The authors applied their experience in the planning, building and running of SOCs to map the similar functions as expressed by Zimmerman and IBM.

Table 1: SOC functions

IBM	Zimmerman
Security and threat monitoring	Real-time monitoring Sensor tuning and management and SOC infrastructure operations and maintenance (O&M) *SOC tool engineering and deployment*
Personnel recruitment, retainment and management	
Process development and optimisation	
Emerging threat strategy (threat intelligence)	*Cyber intelligence collection and analysts*
Security incident management	*Triage or incident analysis, coordination and response*

Cognizance needs to be taken that a SOC function can also be provided as a service depending on the SOC business model. SOCs can be categorised as internal SOCs, or as Managed Security Service Providers (MSSPs). Using the "*Sensor tuning and management and SOC infrastructure operations and maintenance (O&M)"* function in Table II-1 as an example, the following two scenarios serve as illustration:

- Sensor tuning and management and SOC infrastructure O&M need to be performed *as a function* to ensure that all the SOC's infrastructure elements, such as firewalls and IPSs, are maintained and updated. This function is intrinsic to the internal SOC, and is applicable to the SOC as an entity.
- Sensor tuning and management and SOC infrastructure O&M can also be offered by a SOC *as a service* to customers. This is applicable where organisations consume SOC services from MSSPs.

The SOC service delivery models are explained in more detail in Section 4.1.

3.2 CSIRT functions

CSIRT functions as expressed by the Internet Engineering Task Force (IETF) and the SysAdmin, Audit, Networking, and Security (SANS) Institute are listed in Table 2. These were grouped by the authors according to their relevance to each other. This is done to determine the CSIRT functions, as well as illustrate differences and similarities between the IETF and SANS description of CSIRT functions.

Table 2: CSIRT functions

IETF	SANS
Remediate security activity in their constituency	
Play a coordination role to resolve incidents	
Manages the entire incident life-cycle	Provide incident handling capabilities within an organisation
	Real-time incident response activities and non-real-time incident response activities

Following the determination of the SOC and CSIRT high-level functions, the respective management model's services are discussed.

4. SOC and CSIRT services

Although SOCs provide an operational responsibility, an overlap of services does exist when compared with CSIRTs (Kruidhof, 2014). In order to facilitate the completion of the E-CMIRC services integration model, the services offered by both entities have to be determined. This does not imply an analysis or discussion of the service, but is done in order to determine the services applicable and unique to both entities.

The services provided by SOCs are listed below as specified by SOC service providers (Dell, 2013); (DTS Solution, 2015); (HCLTech, 2014); (Hewlett-Packard, 2013); (IBM, 2104); (McAfee, 2013); (Neusoft, 2015); (SecureOps, 2013); (Symantec, 2012); (T-Systems, 2013):

- Counter intelligence.
- Surveillance.

- Integrated threat intelligence.
- Incident response, incident management.
- Asset management and criticality rating.
- Aggregation and analysis of intelligence data.
- Correlation of content intelligence data.
- Analytical intelligence capabilities.
- Workflow automation.
- 24x7 monitoring.
- Forensic analysis.
- In-house research.
- Reporting.
- Vulnerability management.
- Risk management.
- Security knowledge management and security pre-warning.

Services provided by CSIRTs are grouped into three categories according to the European Network and Information Security Agency (ENISA) (2015a). The categories of services, with their related services, are listed below (European Network and Information Security Agency & (ENISA), (2015a) (SEI - Carnegie Mellon, 2002):

- Reactive services - performed in reaction to an incident or request
- *Alerts and warnings.*
- *Incident handling.*
- *Vulnerability handling.*
- *Artefact handling.*
- Proactive services - aids in the preparation, protection and securing of systems in the expectancy of attacks or incidents
- *Announcements.*
- *Technology watch.*
- *Security audits or assessments.*
- *Configuration and maintenance of tools, applications and infrastructure.*
- *Development of security tools.*
- *Intrusion detection services.*
- *Security-related information dissemination.*
- Security quality management services - enhance existing services traditionally performed by other areas of business such as audit and IT
- *Risk analysis.*
- *Business continuity and disaster recovery planning.*
- *Security consulting.*
- *Awareness building.*
- *Education and training.*
- *Product evaluation or certification.*

In order to identify services overlap and similarities, SOC services are mapped to CSIRT services and categories in **Table 3** to **Table 5**. This is done by mapping the SOC services identified in Section 3 to the CSIRT service

categories (reactive, proactive and service quality management service), and services, and by identifying similarities between them.

Table 3: CSIRT to SOC reactive service mapping

CSIRT	SOC
Alerts and warnings	Security knowledge management and security pre-warnings
Incident handling	Incident response and Incident management
Vulnerability handling	Vulnerability Management
Artefact handling	Forensic analysis

Table 4: CSIRT to SOC proactive service mapping

CSIRT	SOC
Announcements	Security knowledge management and security pre-warnings
Technology watch	In-house research
Security audits or assessments	
Configuration and maintenance of tools, applications and infrastructure	Sensor tuning and management and SOC infrastructure O&M
Development of security tools	SOC tool engineering and deployment
Intrusion detection services	24x7 monitoring
Security related information dissemination	Security knowledge management and security pre-warnings

Table 5: CSIRT to SOC security quality management services mapping

CSIRT	SOC
Risk analysis	Risk management
Business continuity and Disaster Recovery Planning for constituents	Part of security consulting
Security consulting	Security consulting
Awareness building	Awareness building
Education and training	Training
Product evaluation and certification	Product assessment (not certification)

Considering the list of proactive services as mapped in Table 4, the service descriptions by vendors (Arcsight, 2009) (ENISA, 2015a) (Kelley et al, 2006) and the definitions of SOCs (McAfee, 2012) and CSIRTs (ENISA, 2015c), it can be inferred that services offered by CSIRTs are mostly focussed towards constituents, while services offered by SOCs are in some cases focussed internally – in other words, the service delivery models differ. Also, the reach differs in that National CSIRTs provide a service to the public as a constituent, while SOCs serve customers, but almost never the public. It can be argued that SOCs do not serve a nation but rather individual clients, or in the case of MSSPs multiple clients. The same applies to CSIRTs in terms of serving multiple constituents. Table 6 lists the similar services, as well as the services unique to SOCs and CSIRTs as defined in Table 3 to Table 5.

Table 6: SOC and CSIRT common and unique services

CSIRT unique	Common	SOC unique
Security audits or assessments against standards	Alerts and warnings Incident handling Vulnerability handling Artefact handling Forensics handling Announcements Technology watch Configuration and maintenance of tools, applications and infrastructure Development of security tools Intrusion detection services Security related information dissemination Risk analysis Business continuity and Disaster Recovery Planning for constituents	Real-time monitoring Reporting and trending Border protection device O&M SOC infrastructure O&M Sensor tuning and maintenance Custom signature creation Insider threat case support and investigation Network mapping Vulnerability scanning and assessment Penetration testing Situational awareness

CSIRT unique	Common	SOC unique
	Security consulting Awareness building Education and training Product evaluation and certification	Media relations

In Section 3 the services common to, and unique to SOCs and CSIRTs were identified. In the next section, the management models as well as scope and reach of SOCs and CSIRTs will be determined. This activity will assist with the development of a services model applicable to the different types of SOCs and CSIRTs (scope) while keeping in consideration their reach.

5. SOC and CSIRT models, scope and reach

In this section, the service delivery model of SOCs and CSIRTs and their respective scopes and reach are discussed. The section concludes with a discussion of common, as well as different technologies applied by SOCs and CSIRTs to facilitate their service delivery.

5.1 Service delivery models

The different service delivery models as well as the scope and reach of SOCs and CSIRTs will have an influence on how the E-CMIRC model is developed. For example, whether a CSIRT serves the public or an organisation internally will influence the services offered by the CSIRT (Kruidhof, 2014). Similarly, the services offered by a SOC will differ if the service is consumed as an outsourced service (as in the case with a MSSP), or offered in-house. The service delivery model of SOCs and CSIRTs in this instance refers to the way in which the service is offered to the consumer of the service. The service delivery model is the way, or structure used to deliver the SOC and CSIRT functions as identified in *Table 1* and *Table 2*. SOC or CSIRT functions differ, but there could in some instances be an overlap in the way the service is delivered, or the structures used to deliver these services.

SOC services are delivered using two distinct models. These are (Rothke, 2012):

- Internal SOCs or SOCs serving the internal security requirements of an organisation. This model serves a single (internal) customer. The deployment could be centralised, or in the case of global organisations, distributed.
- MSSPs providing SOC services-as–a-service to organisations (Gartner, 2013). This model serves different customers form different industries, and even different countries.

SOCs can also be situated locally, or be geographically dispersed (global) (KPMG, 2012). There are advantages and disadvantages for each approach, specifically in terms of cost and skills, but the purpose of this paper is not to analyse these.

ENISA mentions four different ways in which a CSIRT could be structured (ENISA, 2015a). These are:

- As an independent organisation with its own staff (equivalent to a MSSP).
- As part of an existing organisation (equivalent to an internal or local SOC).
- Campus, or distributed model (equivalent to a distributed, or global SOC).
- Voluntary model where people or groups with a shared interest get together and provide advice and support to each other.

Considering the above structures, it is clear that there is little difference between the SOC and CSIRT service delivery models. SOCs typically work in isolation, and it is the experience of the authors that it is not common that information sharing between SOCs and MSSPs take place due to confidentiality clauses signed with customers. CSIRTs are typically further structured into sector CSIRTs which could be constituents of a National CSIRT.

5.2 Scope and reach

In terms of reach, SOCs typically serve a single organisation, whereas MSSPs serve multiple different organisations from different industries. The same statement holds true for CSIRTs, but CSIRTs could also be established to provide a service at both a national and public level. CSIRTs were originally established to provide

incident management services, while incident management is also inherently part of SOCs. It must be kept in mind that the CSIRT services listed in Table 3 - Table 5 will potentially have to be delivered at a national level.

When considering scope and services on offer, it is important to note that services offered exclusively by SOCs are real time monitoring and response, as well as the management of security controls such as firewalls and IPS's. This implies technology additional to what is used in a CSIRT, as well as skills not found in a CSIRT. The list of SOC exclusive services were derived from Table 6.

5.3 Technology

When considering SOCs and CSIRTs from a technology perspective, and excluding common ICT equipment (such as workstations, monitors, phones and video walls since these are common across both SOCs and CSIRTs), one has to look for differences rather than similarities. The purpose of this paper is not to provide an exhaustive list of technologies needed, and only a few samples are provided. Both SOCs and CSIRTs will need the following technologies used to enable their service to constituents (Walker, 2008):

- Call logging system.
- Vulnerability scanners.
- Tool to collect and automate security information dissemination.
- Forensic tools.

A tool unique to SOCs is the Security Incident and Event Management (SIEM) tool to facilitate log and event management (Torres, 2015) (Zimmerman, 2014). Most Commercial off the Shelf (COTS) SIEM's come with built in log management features, correlation rules and automated workflow, as well as a very rudimentary asset management capability.

5.4 Summary

In conclusion, it can be said that the biggest differentiating factor between SOCs and CSIRTs is technology used, and scope and reach. SOCs use a technology specific and unique to their operations – the SIEM. The SIEM is used to receive events from technical controls such as firewalls and intrusion prevention systems, as well as critical assets. These events are managed by the SIEM depending on its capabilities. SIEM capabilities are outside the scope of this paper. These events are also actively monitored by analysts who reacts to them, and elevate events to incidents if necessary. Scope is also a differentiating factor in that SOCs serves customers internally, or provide monitoring as a service to multiple customers. In terms of reach, CSIRTs could provide a service to the population nationally – depending on the type of CSIRT.

6. SOC and CSIRT service integration model

When considering service integration between SOCs and CSIRTs, the most important factors to consider is technology, and scope and reach. As discussed in Section 4.1, SOCs could be internal to organisations, or serve multiple customers when adopting the MSSP model. SOCs can be either local or distributed, while CSIRTs' service delivery models allow for both inter- organisational or national level structures. In addition, the structure can be either local, or distributed.

CSIRTs in South Africa are typically organised into organisational CSIRTs, with sector CSIRTs and a national CSIRT (Minister of Justice and Correctional Services, 2015) (refer to ***Error! Reference source not found.***). Bada *et al* (2014) lists alternative CSIRT organisations applicable to Europe and the United States. These are developed countries, with mature economies and ICT infrastructure, and their organisations will differ from developing countries.

In this model an approach to service integration is proposed on the services applicable to both SOCs and CSIRTs. The model will be developed generically enough to be applied to all service instances. The three criteria which must be met for an organisation to be classified as a CSIRT, and which is also applicable to SOCs are (Zimmerman, 2014) (Shirey, 2014):

- A mechanism for clients or constituents to report cybersecurity incidents.
- A mechanism to provide incident handling assistance to constituents.

- A mechanism to disseminate incident related information to constituents and other third parties.

From Section 3 and Table 6 it can be seen that the incident management capability is central to CSIRT services, and real-time monitoring integral to SOC services. A SOC thus monitors events and elevate them to incidents if needed whereas a CSIRT receives incidents form constituents via telephone, e-mail or any other means, and react to them. A CSIRT never monitors events from technical controls or critical assets and as such does not use SIEM technology. The services as deemed applicable to the proposed E-CMIRC model, is derived from Table 6 and displayed in Table 7. ICT service delivery, and by extension SOC and CSIRT services, has at its core people, processes and technology (ITILnews.com, 2015). The model uses this framework to ensure that all service elements are catered for. The people, process and technology framework underpins the Information Technology Information Library (ITIL) service delivery domain (Leopoldi, 2005).

Table 7: E-CMIRC services

E-CMIRC Services
Incident handling
Real-time monitoring
Alerts and warnings
Intrusion detection services
Security related information dissemination
Infrastructure O&M
Situational awareness

As an operating model, the plan-build-run model is used. Using this next-generation operating model (Claes *et al*, 2014) allows for the breaking down of traditional technology silos and rather focus on ICT functions, thus allowing for growth (Agarwal *et al*, 2013). It is the authors' opinion that following such an approach will provide the best fit for the E-CMIRC at a national level. The people, process and technology framework will ensure that all service requirements for the E-CMIRC are met, while operation of the E-CMIRC is governed by the plan-build-run model.

The E-CMIRC infrastructure needs to be secured, and all governance requirements of the applicable developing country need to be met. These are normally in the form of legal requirements, regulations and others normative documents. The complete high-level model is shown in Figure 1.

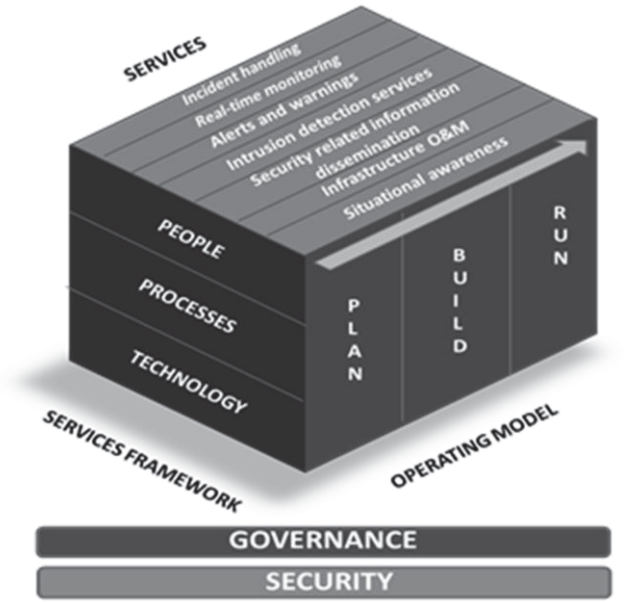


Figure 1: SOC and CSIRT services integration high-level model

7. Future work

In future work, the service selection will be motivated in terms of applicability to requirements of developing countries. Furthermore, the proposed model will be augmented with a mapping of services to the scope and reach of the E-CMIRC, and also identification of applicable standards, best practices and technologies. Maturity

model for the services offered by the E-CMIRC will be developed so as to allow the services to measured, and improved on.
The study will conclude with an identification of infrastructure to be monitored in a developing country, based on E-CMIRC scope and reach. The authors are of the opinion that the completed model will provide for a cost-effective, yet competent, measureable and very powerful initial facility to provide for a cyberdefence capability in developing countries. It will be achieved by using industry accepted best practices and standards, and by combining services traditionally offered by two distinct entities – the SOC and CSIRT.

The number of standards in the model will be limited to avoid an 'alphabet soup' approach. The reasons for the decision are:

- A model based on many standards requires a very large amount of effort and expense to maintain and keep current as the standards involved get updated and change over time.
- Human resources that are familiar with a wide variety of standards are difficult to find and keep.
- If the effort to keep the model up to date is not expended, the model may become less relevant over time.

This approach ensures consistency and ease of maintenance of the model in the future since it is extremely unlikely that international standards bodies will update any particular standard in a way that is inconsistent with the others.

8. Conclusion

Developing countries faces challenges when it comes to securing its infrastructure and national assets. To this effect, a cost effective and competent model for the establishment of an initial capability will go a long way to address these challenges. The model can be used as a baseline for minimum requirements, and as the country matures, additional services can be added to those established during the embryonic phase.

The development of the model started with the identification of SOC and CSIRT functions and services. Different and similar functions and services were then mapped to identify needed services for the model. The scope and reach of SOC s and CSIRTs were described to serve as rationale for the selection of services. This led to the identification of services applicable to the E-CMIRC for use in developing countries. The completeness of the model was ensured using the people, process and technology framework as well as the plan-build-run operational model. This model serves as a reference model for the further development of an initial, or embryonic cybersecurity capability, at national level, for developing countries.

References

Arcsight. (2009). Building a Successful Security Operations Center. Retrieved April 13, 2015, from http://www.scribd.com/doc/39599055/ArcSight-Whitepaper-SuccessfulSOC

Bancroft S. et al. (2013). CISO Information Security Workshop. Retrieved May 18, 2015, from http://digitalstrategies.tuck.dartmouth.edu/cds-uploads/programs/pdf/CISOWorkshop2013final.pdf

C. Zimmerman. (2014). *Ten Strategies of a World-Class Cybersecurity Operations Center*. The Mitre Corporation. Retrieved from https://www.mitre.org/sites/default/files/publications/pr-13-1028-mitre-10-strategies-cyber-ops-center.pdf

Campbell T. (2003). An Introduction to the Computer Security Incident Response Team (CSIRT) Set-Up and Operational Consideration. Retrieved November 24, 2015, from https://www.giac.org/paper/gsec/3907/introduction-computer-security-incident-response/106281

Claes S. et al. (2014). Next Generation IT operating models - KPMG. Retrieved January 5, 2016, from http://www.kpmg.com/BE/en/IssuesAndInsights/ArticlesPublications/Documents/Next-Generation-IT-Delivery-Models.pdf

D. Kelley & R. Morits. (2006). Best Practices for Building a Security Operations Center. *The (ISC)2 Information Systems Security*, *14*(6), 27 – 32. Retrieved from http://www.infosectoday.com/Trial/Kelley.pdf

Danyliw R. (2004). IETF - Incident Object Description Exchange Format (IODEF) Implementation Guide. Retrieved May 26, 2015, from https://www.ietf.org/proceedings/60/I-D/draft-ietf-inch-implement-00.txt

Dell. (2013). Security Operations Centers. Retrieved April 13, 2015, from http://www.secureworks.com/it_security_services/advantage/soc/

DTS Solution. (2015). Security Operations Center 2.0. Retrieved June 1, 2015, from http://www.dts-solution.com/solutions/security-operations-center/

ENISA. (2015a). CSIRT Structure. Retrieved June 4, 2015, from https://www.enisa.europa.eu/activities/cert/support/guide2/internal-management/structure

ENISA. (2015b). Definition of a CSIRT. Retrieved November 24, 2015, from https://www.enisa.europa.eu/activities/cert/support/guide/strategy/what-is-csirt/definition

ENISA. (2015c). What is a CSIRT? Retrieved May 18, 2015, from https://www.enisa.europa.eu/activities/cert/support/guide2/introduction/what-is-csirt

European Network and Information Security Agency, & (ENISA). (2015). CSIRT Services. Retrieved June 1, 2015, from https://www.enisa.europa.eu/activities/cert/support/guide/appendix/csirt-services

Experts Exchange. (2013). Differences between SOC and operation Security team. Retrieved April 13, 2015, from http://www.experts-exchange.com/Security/Misc/Q_28098449.html

Gartner. (2013). Managed Security Service Provider (MSSP). Retrieved November 17, 2015, from http://www.gartner.com/it-glossary/mssp-managed-security-service-provider

Ghernaouti-Hélie S. et al. (2007). *Cybersecurity guide for developing countries*. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwiVu6z8jZfKAhXE2BoKHWQJBZkQFgg4MAM&url=http://www.itu.int/ITU-D/cyb/publications/2007/cgdc-2007-e.pdf&usg=AFQjCNFz7uJc--OKNS6JTP-0tu19dVlQZA&sig2=WyNS

HCLTech. (2014). HCL Managed Security Services -Security Operations Made Simpler. Retrieved June 1, 2015, from http://www.hcltech.com/it-infrastructure-management/managed-security-services

Hewlett-Packard. (2013). 5G/SOC: SOC Generations. Retrieved April 13, 2015, from http://www8.hp.com/us/en/software-solutions/software.html?compURI=1343719#!

IBM. (2016). The 5 essential functions of an enterprise security operations center (SOC). Retrieved May 26, 2015, from http://www-935.ibm.com/services/us/en/it-services/security-services/the-five-essential-functions-of-an-enterprise-security-operations-center-infographic/

IBM. (2104). Virtual Security Operations Center (SOC). Retrieved April 13, 2015, from http://www-935.ibm.com/services/us/en/it-services/virtual-security-operations-center-soc.html

ITILnews.com. (2015). ITIL Back to basics (People, Process and Technology). Retrieved July 20, 2015, from http://www.itilnews.com/ITIL__Back_to_basics_People_Process_and_Technology.html

Jacobs P. (2015). *Towards a framework for building security operation centers*. Rhodes Universtity. Retrieved from http://contentpro.seals.ac.za/iii/cpro/DigitalItemViewPage.external?lang=eng&sp=1017932&sp=T&suite=def

Killcrece G. et al. (2003). State of the Practice of Computer Security Incident Response Teams (CSIRTs). Retrieved January 7, 2016, from http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=6571

KPMG. (2012). Security operations center (SOC) globalization. Retrieved June 4, 2015, from http://www.kpmg.com/SG/en/IssuesAndInsights/ArticlesPublications/Documents/Advisory-CS-Security-Operations-Center-SOC-Globalization.pdf

Kruidhof O. (2014). Evolution of National and Corporate CERTs - Trust, the Key Factor. doi:10.3233/978-1-61499-372-8-81

Leopoldi R. (2005). The Role of ITIL in IT Governance - ITSM. Retrieved January 5, 2016, from www.itsm.info/ITSM ITIL and IT Governance.pdf

M. Bada et al. (2014). Computer Security Incident Response Teams (CSIRTs) An Overview. Retrieved June 5, 2015, from https://webcache.googleusercontent.com/search?q=cache:JudMhFWvjTwJ:https://www.sbs.ox.ac.uk/cybersecurity-capacity/system/files/CSIRTs.pdf+&cd=1&hl=en&ct=clnk

McAfee. (2012). Focus on 5 SIEM Requirements. Retrieved April 13, 2015, from http://www.mcafee.com/us/resources/brochures/br-focus-on-five-siem-requirements.pdf

McAfee. (2013). Case Study McAfee's Unique Prevent-Detect-Respond Approach and Security Operations Center Showcase Best Practices. Retrieved from http://www.mcafee.com/us/resources/case-studies/cs-mcafee-inc.pdf

Minister of Justice and Correctional Services. (2015). South African Cybercrimes and Cybersecurity Bill - Draft for Public Comment. Retrieved November 26, 2015, from http://www.justice.gov.za/legislation/invitations/CyberCrimesBill2015.pdf

NATO Cooperative Cyber Defence Centre of Excellence. (2012). *National Cyber Security Framework Manual*. Retrieved from https://ccdcoe.org/publications/books/NationalCyberSecurityFrameworkManual.pdf

Neusoft. (2015). NetEye Security Operations Center (SOC). Retrieved June 1, 2015, from http://www.neusoft.com/products&platform/1338/

Rothke, B. (2012). Building a Security Operations Center (SOC). Retrieved June 4, 2014, from http://www.rsaconference.com/writable/presentations/file_upload/tech-203.pdf

SecureOps. (2013). SecureOps Security Operations Center. Retrieved April 13, 2015, from http://secureops.com/Services/Monitoring-Services.html

SEI - Carnegie Mellon. (2002). CSIRT Services. Retrieved December 12, 2015, from http://www.cert.org/csirts/services.html

SEI at Carnegie Mellon University. (2015). CERT Division Frequently Asked Questions (FAQ). Retrieved May 18, 2015, from http://www.cert.org/faq/

Shirey R. (2014). Internet Security Glossary, Version 2. *IETF RFC 4949*. Retrieved June 5, 2015, from https://tools.ietf.org/html/rfc4949

Symantec. (2012). Symantec Unveils New Global Security Operations Center in U.S. Retrieved April 13, 2015, from http://www.symantec.com/about/news/release/article.jsp?prid=20120207_01

Tagert A. (2010). *Cybersecurity Challenges in Developing Nations*. Carnegie Mellon University. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0ahUKEwiVu6z8jZfKAhXE2BoKHWQJBZkQFggwMAI&url=http://repository.cmu.edu/cgi/viewcontent.cgi?article=1021&context=dissertations&usg=AFQjCNHC4TZyrO9d-zFjzUXm5Z8IU7w82w&sig2

Torres A. (2015). Building a World-Class Security Operations Center: A Roadmap. Retrieved November 17, 2015, from https://www.sans.org/reading-room/whitepapers/analyst/building-world-class-security-operations-center-roadmap-35907

T-Systems. (2013). Security. Retrieved from http://www.t-systems.com/cebit/cyber-defense-strategies-and-secure-mobile-communication-by-t-systems/1033478

Walker T. (2008). Practical management of malicious insider threat – An enterprise CSIRT perspective. *Information Security Technical Report*, *13*(4), 225–234. doi:10.1016/j.istr.2008.10.013

Weinberg, A., Agarwai, N., & Bommadervara, N. (2013). Using a plan-build-run organizational model to drive IT infrastructure objectives. *McKinsey & Company Website*. Retrieved January 5, 2016, from https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=Plan Build Run

A Generic Framework for Digital Evidence Traceability

Nickson Karie, Victor Kebande and Hein Venter
Department of Computer Science, University of Pretoria, South Africa
menza06@hotmail.com
vickkebande@gmail.com
heinventer@gmail.com

Abstract: Digital forensics has emerged as a discipline that plays a very critical role in both civil and criminal cases. However, due to the evolution in digital technologies, the recovery and investigation of evidence found in digital devices, often in relation to digital crimes, is increasingly becoming sophisticated. Digital forensic investigators, unluckily, can only work with existing digital forensic investigation tools to locate, gather, analyse and reconstruct data from computer systems, networks, wireless communications and other digital devices. Unfortunately, most of the existing digital forensic investigation tools consist of dissimilar elements or parts and are consequently unable to work together harmoniously. The ability to chronologically interrelate uniquely identified digital forensic evidence data from a crime scene with tools that are unable to work together harmoniously makes digital evidence traceability a challenge to quite a number of investigators. The aim of this paper, therefore, is to propose a generic framework for digital evidence traceability. Such a framework is meant to assist or guide digital forensic investigators, law enforcement agencies and other digital forensic practitioners in identifying, tracking or even verifying the source and history, including the application of specific digital evidence data captured during an investigation process. In the authors' opinion, employing such a framework in digital forensics can help investigators to save time as well as simplify the digital evidence traceability process.

Keywords: digital forensics, generic framework, digital evidence, traceability

1. Introduction

When considering the digital forensic investigation process, the knowledge, skills, tools and techniques used to access, collect, analyse and organise Potential Digital Evidence (PDE) from computing systems, networks, wireless communications, and storage devices are critically becoming indispensable. Moreover, to convince the court that the gathered PDE is worthy of inclusion into the criminal process, investigators must use extensive technical knowledge and skills, including tools and techniques that are typically designed for handling digital evidence (Karie and Venter, 2013). Unfortunately, most of the existing Digital Forensic Investigation (DFI) tools consist of dissimilar elements or parts and are consequently unable to work together harmoniously. Accordingly, the ability to chronologically interrelate uniquely identified digital evidence from crime scenes with tools that are unable to work together makes PDE traceability a challenge to digital forensic investigators.

According to Arnold and Soriano, (2013) and Roberts and Suits, (2013) the admissibility of PDE in any court of law is nowadays coming under increased scrutiny. This implies that, the failure to correctly identify and verify the source, history and/or the application of the captured digital evidence during an investigation process can make it hard for such evidence to be considered for inclusion in the legal argument. Further to this, the investigators, Law Enforcement Agencies (LEA) and other digital forensic practitioners equally expect that any potential evidence captured during an investigation process must remain unaltered in order for it to be suitable for litigation. This means that, the identity, history and the exact source or origin of captured PDE must be verifiable through the existing forms of legal argument and scientifically accepted methods.

The aim of this paper, therefore, is to propose a generic framework for digital evidence traceability in digital forensics. Note that, traceability is used in this paper to connote the ability to identify and verify the source, history and the application of PDE data captured during digital investigation process, which also includes the audit trails and evidence assurance. The proposed framework in this paper is meant to assist investigators, LEAs and other digital forensic practitioners in identifying, tracking and verifying the source, history, and the application of specific captured PDE during an investigation process. By using such a framework for the identification and verification of the captured PDE as well as the audit trail, in the authors' opinion, can save time and simplify the traceability process.

As for the remaining part of this paper, section 2 briefly presents background concepts on digital forensics (DF), evidence traceability and PDE. Section 3, on the other hand, concentrates on discussing related work. An explanation of the proposed framework is handled in section 4 followed by a discussion of the framework in section 5. Finally, conclusion and future work is given in section 6.

2. Background

This section introduces the reader to the background study on the following parameters: Digital forensics, evidence traceability and PDE. Digital forensics is discussed to show the essence of post-event response during DFI while evidence traceability is explained to show the usefulness of identifying and verifying the source and history of captured PDE during a DFI process. Finally, PDE is discussed to show the role the recovered digital artefacts plays during hypothesis creation in a court of law for civil or criminal proceedings.

2.1 Digital forensics

Digital forensics (DF) is considered the branch of forensic science dealing with the recovery and investigation of material found in digital devices, often in relation to digital crimes. Concepts of DF were first realised in Utica, New York during the first Digital Forensic Research Workshop (DFRWS) in 2001. During this time DF was defined as the use of proven and scientifically derived methods toward collection, identification, validation of artifacts that represents digital evidence which may further lead to reconstruction of events that may be deemed as criminal (Palmer, 2001). This definition represents an investigative process for digital forensic science since science and technology is involved in developing and testing of theories.

Since its inception DF has, however, continued to penetrate the society. This has made both individuals and organisations today to sense the different opportunities that have come with this new discipline. However, the techniques employed in modern crime are also advancing each day; hence the need for standardized incident-response techniques that may help to solve the ever-rising cyber threats and attacks. When considering the DFI process, the captured PDE is arguably one among the most significant artefacts of the process (Yusoff, et al. 2011).

For this reason, the investigators involved should be competent and proficient in all the investigation processes used (Karie and venter, 2013). Further, the investigation processes should be compatible with the relevant policies and/or laws in various jurisdictions. Additionally, the procedures and techniques used in the DFI process should also allow for digital evidence traceability so as to able to be admitted to a court of law (Carrier, 2006). Despite that, if digital evidence is not properly or legally acquired it may not be admissible in a court of law. In conclusion, a traceability report can be used to shade more light on the source and history of any captured evidence. The next section elaborates on evidence traceability in DF.

2.2 Evidence traceability

Traceability is the ability to verify the source, history, location, or application of an item by means of documented or recorded identification (Rupali and Tabassum, 2012). In today's world the traceability process has become an important component of many parts of our lives including the digital investigation process. This is backed up by Siti et al. (2011) where the authors state that, traceability is capable of mapping events of an incident from difference sources and as a result help in obtaining evidence of an incident to be used in a court of law or any other investigation aspects. The main goal of evidence traceability is usually to discover PDE and connect meaningful relationships between the discovered evidence. Given the origin of an object, traceability provides the opportunity to track a chain of events as well as predict process outcomes (Siti et al., 2013). This therefore, helps investigators and LEAs to trace all the aspects of any digital evidence based on its origin and history. The next section explains in brief the concept of PDE.

2.3 Potential digital evidence

Carrier and Spafford (2004) have presented Digital Evidence (DE) as an object. Obviously, being an object it has unique features and characteristics based on their functionality and their creator. Apart from that, the notion that digital data has a physical form implies that physical evidence may contain digital evidence. Discounting that, DE is an aspect that cannot be overlooked because it may not be obvious when a computer-based crime occurs. Due to this, an efficient framework that has a capability of providing pertinent DE traceability can ensure ease of detection of potential security incident by digital forensic analysts as well as finding original sources and history that caused the incident (Selamat, 2011).

On the same note, the legal considerations for admissibility of DE vary across diverse jurisdictions. For example, the legality and reliability of DE will be considered as a requirement when deciding on the admissibility of DE

(Hatch, 2008). It is for such reasons that Vacca (2005) highlighted that the general forensic procedures that have to be followed while investigating DE include but are not limited to the following: identification, preservation, acquisition, authentication and analysis.

With the advances in digital technology as stated by Karie and venter (2014) the sources of digital evidence have also grown exponentially. For this reason, it is important that investigators and LEAs establish reliable sources for each of the different types of digital evidence captured during the investigation process (Karie and venter, 2014). Failure to establish reliable sources of any captured digital evidence, for example, can make it hard for investigators to produce a reliable digital evidence traceability report. This is also backed up by the fact that, having unreliable digital evidence sources can potentially be more damaging than having no sources at all (Karie and venter, 2014). Therefore, when gathering PDE, the use of scientifically-proven methods is recommended. Such methods, however, when used in digital forensics must be based on empirical and measurable standards subject to specific scientific principles (Karie and venter, 2014). The next section handles related work in this paper.

3. Related work

This section presents different researcher's relevant works that is somewhat used as related works in this paper. Mohammed and Manaf (2014) argue that in digital forensics, traceability has a significant role during a digital investigation process. This is mainly because it is used in the process of identifying the chain of evidence. The authors in their paper proposed an enhancement of a traceability model that is based on the scenarios for DFI process. Their proposed model consists of stakeholders, object and source phases that are used to find the relationship between evidence.

A research paper by Kebande and Venter (2015a) as well as Kebande and Venter (2015b) the authors highlighted a model for characterizing PDE. In their study, the authors were able to propose a model that was able to characterize PDE based on the causality and characteristics of evidential activities during pre-analysis stage of evidence collection. Even though the studies main focus was primarily on the cloud environment and forensic readiness, the study also portrayed a relationship with traceability with regard to how digital evidence can be identified and verified. On the same note, Selamat et al., (2011) presents traceability as a process of mapping and tracing digital evidence so that the source of a potential security incident can be established. This can only be possible by trying to establish a trace pattern from protocols, open ports, networks, internet activities, cloud environment and electronic devices to enable the digital forensic investigator to find the source of the security incident.

An experimental design presented by Selamat et al (2013) to establish the trace and map concepts in traceability in a digital forensic perspective had the following components: Inquisition of incident scenario, identification of incident trace pattern and a construction of tracing and mapping pattern. In their study the tracing rate, mapping rate and offender identification rate was used to present the level of tracing ability, mapping ability and offender ability respectively. Lastly a study by Kebande and Venter (2014) on sources of PDE capture portrays a way through which sources of evidence can be captured from different scenarios. Even though the authors never employed traceability techniques they portrayed different ways through which PDE can be traced and captured using an agent-based solution presented in there paper as a Non-Malicious Botnet (NMB).

There exist other related works on issues similar to digital evidence traceability, however, neither those nor the cited references in this paper have presented a generic framework to assist investigators in PDE traceability in the way that is introduced in this paper. However, the authors acknowledge the fact that the previous research works have offered useful insights toward the development of the proposed generic framework in this paper. The next section introduces the reader to the proposed framework.

4. The proposed framework for potential digital evidence traceability

In this section of the paper, the authors present a detailed explanation of the proposed framework as a new contribution to this study. Figure 1 shows the structure of the proposed framework for digital evidence traceability.

The framework consists of three steps arranged from top to bottom where Step 1 (labelled: Input) accepts the captured and unaltered PDE as input. This is followed by a series of processes in Step 2 (labelled: Processes) that

works on the captured PDE. The alphabets A, B and C that are shown in Step 2 represent the aforementioned processes. Finally Step 3 (labelled: Output) supplies a variety of reports in the form of output from the proposed framework.

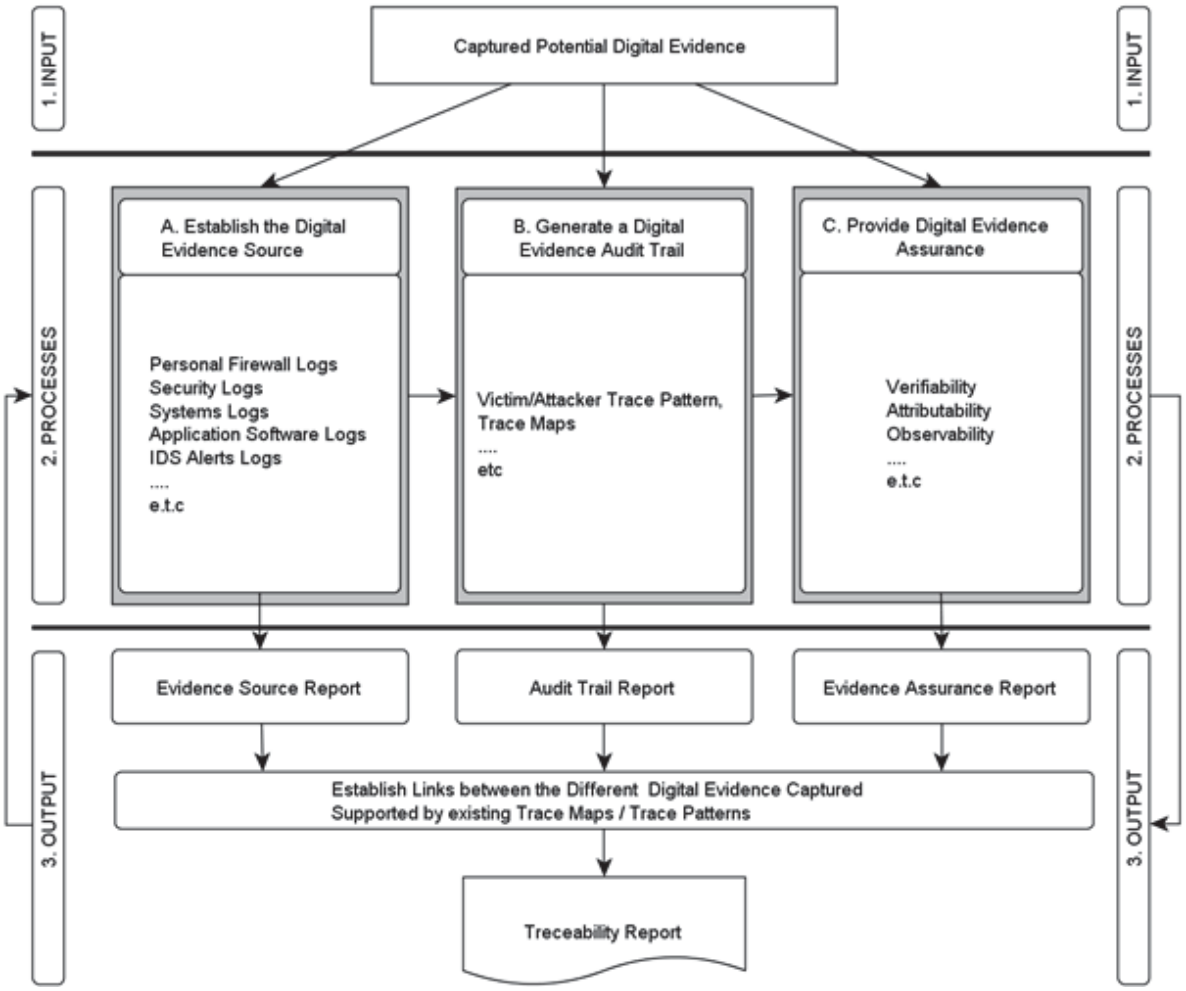


Figure 1: Proposed framework for digital evidence traceability

Note that the authors throughout this paper uses the term 'potential' digital evidence because according to Karie and Venter (2013) digital artefacts are only considered to be 'evidence' in the final phase of the digital forensic investigation process, namely the reporting phase. This also implies that, for the captured PDE to be considered as competent digital evidence according to Ryan and Shpantzer (2005), it must possess scientific validity grounded in scientific methods and procedures as described in the international standard ISO/IEC 27043: 2015. In the sections to follow, the Steps labelled 1 to 3 as shown in the proposed framework in Figure 1 are further explained.

4.1 Input

In digital forensics, there exists a variety of PDE that can be captured using different forensic tools. As mentioned by Karie and Venter (2013) the requirement for PDE presupposes that all forms of digital evidence should be considered. Such evidence may include, but are not limited to: log files, emails, images, video clips, electronic documents, back-up disks, portable computers, network traffic records, personnel records, access control systems and telephone records.

However, before using any of such PDE to determine the truth of an issue, the investigator must be sure that the evidence has been captured and its origin or source can be established and verified with a high degree of certainty. This is backed up by the fact that, the admissibility of any captured PDE in any court or legal proceedings is further subject to examination and verification through existing forms of legal argument. However, having proof of originality of all the primary sources of the captured PDE is essential in coming up with a reliable traceability report. Furthermore, the proof of originality of all captured evidence can be useful

especially on the inferences drawn from such PDE. The concept of establishing the digital evidence source is, however, discussed in Step 2 of the proposed framework. Although it is beyond the scope of this paper to further elaborate on the individual types of PDE that can be captured, the knowledge and proof of its primary origin is essential in coming up with reliable digital evidence traceability reports. The next section will explain the different process as shown in Figure 1.

4.2 Processes

For a comprehensive digital traceability report, different processes may be necessary. In this paper, however, the most general processes are discussed. The investigators are, however, free to employ additional processes as deemed fit other than those mentioned in the framework. This is because the digital technology environment is very dynamic and new ideas always emerge, hence, the processes explained in this section of the paper are by no means the only final processes to a good evidence traceability report. The processes are: Establish the digital evidence source, generate a digital evidence audit trail and finally, provide digital evidence assurance. These processes are explained in the subsections to follow.

4.2.1 Establish the digital evidence source

According to Karie and Venter (2013), it is important that investigators identify reliable sources of each of the different types of PDE captured during an investigation process before using it in court. The sources may include: personal firewall logs, security logs, systems logs, application software logs, IDS alerts logs among others as shown in Figure 1. Note also that, the PDE sources may exist in different forms (primary or secondary form). Primary sources are usually first-hand sources; for example, photographs captured using a digital camera, from e-mail or from recorded speeches. Secondary sources also referred to as second-hand sources may include: information distributed freely online or information on printed materials.

However, a secondary source may also be a primary source depending on how it is used (Ithaca, 2013). This is backed up by the fact that, "primary" and "secondary" are relative terms and, therefore, sources can be judged as primary or secondary depending on their specific contexts and according to what they are used for (Helge, 1989). Therefore, the digital forensic experts should be well versed with the exact type of evidence at hand, the exact source; either primary or secondary as explained above, and where such digital evidence was captured. Failure to identify the source of the PDE, for example, can make it hard for such digital evidence to be considered for inclusion in any legal argument (Karie and Venter, 2013). Once the digital evidence source has been established, this process leads to the generation of an audit trail as shown in Figure 1 and is discussed in the next subsection.

4.2.2 Generate a digital evidence audit trail

An audit trail can be viewed as a documentation or set of records that can be used to show who accessed a particular computer system during a given period of time and what events or procedures and operations he or she performed during a particular run. However, according to NIA (2016), an audit trail (also called audit log) can be defined as “a security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, or event”.

Knowing that the goal of digital forensics is to examine digital media in a forensically sound manner with additional guidelines and trusted procedures designed to create legal audit trails, the proof of clear and original audit trails, thus, plays an important role in digital forensics (Karie and Venter, 2015), more especially in evidence traceability. Audit trails are useful both for maintaining security and for recovering lost transactions as they can show the victim/attacker trace pattern as shown in Figure 1. However, it is possible that an intruder may edit or delete the audit trail on a computer, especially weakly protected personal computers (Yong, 2013). Sophisticated rootkits that dynamically modify kernels of running systems to hide what is happening, or even to produce false results are also on the increase. As shown in Figure 1, once the audit trail is generated, an investigator can then proceed to provide a digital evidence assurance as discussed in the next subsection.

4.2.3 Digital evidence assurance

Rob et al. (2005) defines assurance as a qualitative statement expressing the degree of confidence that a claim is true. It can also be defined as a positive declaration intended to give confidence on an important topic or

matter in question. In the case of this paper the digital evidence assurance is meant to offer a positive declaration intended to give confidence on the verifiability, attributability and observability of the source of the PDE captured during an investigation process as well as the evidence audit trail as shown in Figure 1. This also implies that, the digital evidence assurance in this paper is meant to give a positive declaration intended to give confidence on the accuracy and reliability of the evidence source as well as the evidence audit trail which were discussed earlier in section 4.2.1 and section 4.2.2 respectively.

Note that, digital evidence verifiability as shown in the proposed framework is meant to prove the truth of the potential evidence presented in a court of law or civil proceedings. Attributability on the other hand is meant to show the degree to which the PDE captured can be attributed to an individual, an event or even a process. Observability is a measure that can be used to show how well the PDE can be inferred through the knowledge of its attributes including the trace patterns and trace maps. The end result of all the processes in Step 2 are: evidence source report, audit trail report and evidence assurance report as shown in Figure 1. Step 3 of the proposed framework forms part of the output and is explained in the next section.

4.3 Output

The evidence source report, the audit trail report and the evidence assurance report form the output part of the proposed framework. Note also from Figure 1 that the processes (Step 2) and the output (Step 3) allow one to move back and forth until the desired and reliable traceability report is achieved. This implies that there is no limitation on the reiteration one can make on these steps. The evidence source report, audit trail report and the evidence assurance report as shown in Figure 1 together are used to establish the links between the different digital evidence captured as supported by existing trace maps / trace patterns which finally produces the traceability report. Given the origin of any captured digital evidence, traceability can then provide the opportunity to trace a chain of events, or to predict process outcomes. Besides, as defined by (Siti et al., 2013) traceability is the ability to trace and map the events of an incident from different sources in order to obtain useful evidence. The process of generating trace maps can then be made easier with the availability of the victim/attacker trace patterns which helps in discovering the origin or starting point of an incidence that has happened. A reliable traceability report is what the proposed framework in this paper aims to achieve. The next section will present a brief discussion of the proposed framework.

5. Discussion of the proposed framework

The proposed framework in this paper is a new contribution in the digital forensics domain. The scope of the framework is defined by the steps shown Figure 1. The main steps as depicted in the framework include:

- Input (Step 1)
- Process (Step 2)
- Output (Step 3)

The specific details of the individual steps as identified in the framework have further been explained in this paper. However, note that the steps as identified in Figure 1 are meant to facilitate this study and primarily focus on PDE traceability. Such proposed steps are by no means the final guaranteed steps to accurate digital evidence traceability reports. This therefore, implies that, any processes used to manage all the three steps (Step 1 to Step 3) must possess scientific validity grounded in scientific methods and procedures. In the authors' opinion, however, organising the framework into steps was necessary to simplify the understanding of the framework as well as to present specific finer details of the framework.

The proposed framework in this paper can be used in the digital forensics domain, for example, to help investigators in coming up with trace maps based on the digital evidence captured as well as existing audit trails. The framework can also be useful in identifying relevant attack patterns and events to be incorporated in a digital evidence traceability report. Moreover, the digital evidence traceability report drawn from the framework can also be helpful to law enforcement agencies and other stakeholders, for example, in reasoning and identifying specific digital evidence that is relevant to support or refute a particular criminal case in court. For the case of digital evidence admissibility in legal proceedings, the steps as identified in the framework can be used, for example, to evaluate the source, history, validity and reliability of the digital evidence captured.

In addition, developers of digital forensics tools can also use the proposed framework to develop digital evidence traceability tools that can automate the process of creating trace maps and attackers trace patterns. This also implies that developers might find the framework in this paper useful, especially when considering the development of new digital forensic tools and techniques for addressing digital evidence traceability. Finally, the framework presented in this paper has been designed in such a way as to accommodate new processes that may emerge as a result of legal requirements or domain change. To the best of the authors' knowledge, this is a new contribution towards advancing the digital forensic domain.

6. Conclusion

If we revisit the problem addressed in this paper, we are able to coin it as follows: Is it possible to propose a framework for digital evidence traceability in digital forensics that can guide LEAs, DF practitioners and digital forensic investigators? The authors have answered this by proposing a traceability framework that has been discussed in Section 4. There are currently not many standardised guidelines that have specifically been designed to help in digital evidence traceability in digital forensics. The proposed framework in this paper, therefore, is an attempt to provide a way to approach digital evidence traceability. The requirement of such a framework in digital forensics is exceptionally important to investigators, especially during legal hearings in court or civil proceedings. With such a framework, investigators will, for example, be able to structure traceability reports as well as identify relevant trace patterns of evidence to be incorporated during the presentation of digital evidence reports in court. Moreover, the framework can also help law enforcement agencies, as well as the jury in evaluating opinions that substantially outweighs prejudicial effect. Finally, the authors believe that by using such a framework, digital evidence traceability can substantially be attained. However, more research needs to be conducted in order to improve on the proposed framework in this paper. The framework should also spark further discussion on the development of new techniques to support digital evidence traceability in digital forensics.

References

Arnold, E., and Soriano, E., (2013).The Recent Evolution of Expert Evidence in Selected Common LawJurisdictions Around the World. A commissioned study for the Canadian Institute of Chartered Business Valuators.

Carrier, B., & Spafford, E. H. (2004, July). An event-based digital forensic investigation framework. In *Digital forensic research workshop* (pp. 11-13).

Carrier, B.D., (2006). Digital Investigation and Digital Forensic Basics. Available at: http://www.digital-evidence.org/di_basics.html [Accessed March 28, 2013].

Gary Palmer, A Road Map for Digital Forensic Research. Technical Report DTR-T0010-01, DFRWS, November 2001. Report from the First Digital Forensic Research Workshop (DFRWS).

Hatch, B. 2008. *Hacking exposed linux.* 3rd edition. Emeryville, CA: McGraw Hill Osborne Media

Helge, K., (1989). An Introduction to the Historiography of Science. Cambridge University Press. p. 121. ISBN 0-521-38921-6.

ISO/IEC 27043: 2015. Information technology - security techniques - digital evidence investigation principles and processes. http://www.iso27001security.com/html/27043.html (Accessed March 17,2016).

Ithaca College Library, (2013). Primary and secondary sources. Available at: http://www.ithacalibrary.com/sp/subjects/primary [Accessed April 1, 2013].

Karie, N.M. & Venter, H.S., (2015). Taxonomy of challenges for Digital Forensic Disciplines. Journal of Forensic Sciences. Vol. 60, No. 4: 885-893. Online ISSN: 1556-4029

Karie, N.M. and Venter, H.S., (2013).Towards a Framework for Enhancing Potential Digital Evidence Presentation. In the Proceedings of the 12th Annual Information Security for South Africa Conference, (ISSA-2013). Johannesburg, South Africa. Published online by IEEE Xplore®.

Karie, N.M. and Venter, H.S., (2014). A Generic Framework for Enhancing the Quality of Digital Evidence Reports,. In proceedings of the 13th European Conference on Cyber Warfare and Security (ECCWS-2014), Piraeus, Greece.

Kebande, V. R., & Venter, H. S. (2014). A Cloud Forensic Readiness Model Using a Botnet as a Service. In The International Conference on Digital Security and Forensics (DigitalSec2014) (pp. 23-32). The Society of Digital Information and Wireless Communication.

Kebande, V. R., & Venter, H. S. (2015a). Obfuscating a Cloud-Based Botnet Towards Digital Forensic Readiness. In *Iccws 2015-The Proceedings of the 10th International Conference on Cyber Warfare and Security* (p. 434). Academic Conferences Limited.

Kebande, V., & Venter, H. (2015b). Towards a Model for Characterizing Potential Digital Evidence in the Cloud Environment during Digital Forensic Readiness Process. In *ICCSM2015-3rd International Conference on Cloud Security and Management: ICCSM2015* (p. 151). Academic Conferences and publishing limited.

Mohamed, I. A., & Manaf, A. B. A. (2014, April). An enhancement of traceability model based-on scenario for digital forensic investigation process. In 2014 Third International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec).

NIA (National Information Assurance) Glossary (2016)". Committee on National Security Systems. Available at: http://www.icaiknowledgegateway.org/littledms/folder1/glossery.pdf [Accessed February 13, 2016].

Rob W., Georgios D.,Tim K.,and John M. (2005). Combining Software Evidence–Arguments and Assurance. Proceedings of the 2005 workshop on Realising evidence-based software engineering. ACM 1-59593-121-X.

Roberts, J.L, and Suits, C., (2013). Admissibility of digital image data: concerns in the courtroom. Available at: http://libraries.maine.edu/Spatial/gisweb/spatdb/acsm95/ac95071.html [Accessed April 10, 2013].

Rupali B. and Tabassum K. (2012).Tracebility in Digital Forensic Investigation Process. International Journal of Advanced Research in Computer Science and Software Engineering. Vol 2, No. 10

Ryan, D.J., and Shpantzer, G., (2005). Legal Aspects of Digital Forensics. Available at: http://euro.ecom.cmu.edu/program/law/08-732/Evidence/RyanShpantzer.pdf [Accessed April 1, 2013].

Selamat, S. R., Sahib, S., Hafeizah, N., Yusof, R., & Abdollah, M. F. (2013). A Forensic Traceability Index in Digital Forensic Investigation.

Selamat, S. R., Yusof, R., Sahib, S., Roslan, I., Abdollah, M. F., and Mas' ud, M. Z. Adapting Traceability in Digital ForensicInvestigation Process. Malaysian Technical Universities International Conference on Engineering & Technology (MUiCET 2011), 2-3; 2011.

Siti R.S., Robiah Y., Shahrin S., Irda R., Mohd F.A., Zaki M. (2011) Adapting Traceability in Digital Forensic Investigation Process. Procedings of the Malaysian Technical Universities International Conference on Engineering & Technology (MUiCET 2011) pp. 461-468s

Siti R.S., Shahrin S., Nor H., Robiah Y., Mohd F.A. (2013)A Forensic Traceability Index in Digital Forensic Investigation. Journal of Information Security. pp 19-32.

Vacca, J. 2005. Computer forensics – Computer crime scene investigation. 2nd edition. Hingham: Charles River Media.

Yong G. Digital forensics: research challenges and open problems. http://itsecurity.uiowa.edu/security day/documents/guan.pdf (accessed June 21, 2013).

Yusoff, Y., Ismail, R. and Hassan, Z., (2011). Common Phases of Computer Forensics Investigation Models. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No. 3.ISO/IEC 27043: 2015 Information Technology-Security Techniques Incident Investigation Principles and Processes, [online], Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=44407 (Accessed April 17,2016).

Towards a Prototype for Achieving Digital Forensic Readiness in the Cloud Using a Distributed NMB Solution

Victor Kebande, Hermann Stephane Ntsamo and H.S.Venter
Information and Computer Architecture Research Lab (ICSA), Department of Computer Science, University of Pretoria, South Africa
vickkebande@gmail.com
hermann.stephane.ntsamo.work@gmail.com
hventer@cs.up.ac.za

Abstract: Our implementation is aimed at proving the possibility of achieving Digital Forensic Readiness in (DFR) the cloud environment without having to modify the functionality of existing cloud architecture. Considering the distributed and the elasticity of the cloud, there lacks an easy way of conducting DFR without employing a novel software application as a prototype. In this paper therefore, the authors have developed a software application with a functionality of a modified form of a botnet that is able to collect digital information that exist as Potential Digital Evidence (PDE) from the cloud environment, digitally preserve it and store it in a forensic database for forensic readiness purposes. The experiments conducted in this paper have shown promising results because integrity of collected digital information and the legal aspects has been maintained at the same time. Nevertheless, the importance of such a prototype include maximizing the use of PDE while reducing the time and cost needed to perform a Digital Forensic Investigation (DFI). The guidelines that have been used while conducting this process complies with the standard of ISO/IEC 27043: 2015. Based on this premise the authors are able to prove that DFR can be achieved in the cloud environment using the novel software prototype.

Keywords: cloud, forensic; readiness, prototype potential, digital, evidence, botnet, NMB, solution

1. Introduction

Device proliferation has forced organizations and institutions' infrastructures to be built on cloud-based infrastructures which has further led to integration and development of a foundation through which a single application can be provisioned to multiple users at a lesser cost. This has eased the way through which IT services are deployed in larger organizations through efficient communication, quality of services and simplified information-centric methods of delivering solutions.

Discounting that, the distribution and elasticity of the cloud has, in this regard, made it a challenge to conduct Digital Forensic investigation(DFI) because of the costs associated with the modification and tampering of the functionality of the existing cloud architecture(Kebande & Venter, 2014a). The importance of Digital Forensic Readiness (DFR) while conducting these investigations according to Rowlingson (2004) will be to maximize the potential use of DF evidence while minimizing the cost of performing a DFI. At the time of writing this paper, there still lacks an easy way of conducting DFR in the cloud without having to modify the functionality of the existing architecture. Nevertheless, there still exist no accepted techniques of conducting DFR in the cloud apart from the ISO/IEC 27043: 2015 international standard, however, ISO/IEC 27043 is not focused in the cloud environment. In spite of that, there also exist no software prototypes that will help to achieve DFR in the cloud using a Non-Malicious Botnet (NMB) as a forensic agent. The implementation conducted in this paper complies with readiness processes that have been defined in the ISO/IEC 27043: 2015.

The remainder of the paper is structured as follows: Section 2 provides preliminary work on digital forensics, cloud computing and the botnets. Part of this work has been proposed by the authors in their previous work. After that, Section 3 presents related work which is then followed by Section 4 that presents the proposed prototype implementation. Section 5 concentrates on discussing the legal and compliance aspects on admissibility of digital evidence. Thereafter Section 6 provides a critical evaluation of the paper while Section 7 concludes this paper and gives indications of future work.

2. Preliminary work

In this section a discussion on the following parameters is presented as preliminary work: Cloud forensic readiness (CFR) and distributed forensic client. CFR is discussed to show the benefits of reactive and proactive forensics in the cloud environment. Botnets are presented as agents whose functionality is modified to collect digital forensic information.

2.1 Cloud forensic readiness: An overview

Cloud Forensic Readiness (CFR) is a mechanism that allows the addition of Digital Forensic Readiness (DFR) to the cloud environment**.** However, based on the guidelines for developing standards provided by National Institute of Standards and Technology (NIST), the cloud has been characterized from a taxonomy based on very important aspects that are concerned with deployment strategies. The cloud has been represented as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell & Grance, 2011).

On the same note, DFR according to Rowlingson (2004) has an aim of maximizing the use of Potential Digital Evidence (PDE) while reducing the cost and time needed to conduct a Digital Forensic Investigation (DFI). DFR in the cloud can therefore be implemented through active collection of digital information that is related to crimes. The process of proactive preparation before potential security incidents can occur has been presented based on a business perspective where active collection of PDE in form of log files, back-up files, emails and traffic records are collected before a potential crime (Rowlingson, 2004). An important aspect of enforcing this strategy is to help a given organization to be able to manage any availability of business risks because the presence of PDE can be able to create a hypothesis that can support a legal defense in a court of law.

2.2 Distributed forensic client solution

The principles behind using a distributed forensic client solution as an agent in the cloud is to modify the nefarious functionality of the botnet infrastructure in order to rally the botnet to attempt to "infect" virtual instances in the cloud environment through the Command and Control(C&C) server with a positive connotation. The NMB solution is able to proactively collect digital information in a forensic readiness manner that may be used as digital evidence. The NMB solution operator is able to install the NMB infection vectors into a client's infrastructure in the cloud and additionally the service is offered through an operation of a dropping zone for the digital evidence that has been captured. The process is done according to the ISO/IEC 27043: 2015 guidelines on forensic readiness process. Potential Digital Evidence (PDE) integrity and integrity verification is an important aspect during this process of forensic evidence capture. Figure 1 shows a block diagram of the process.

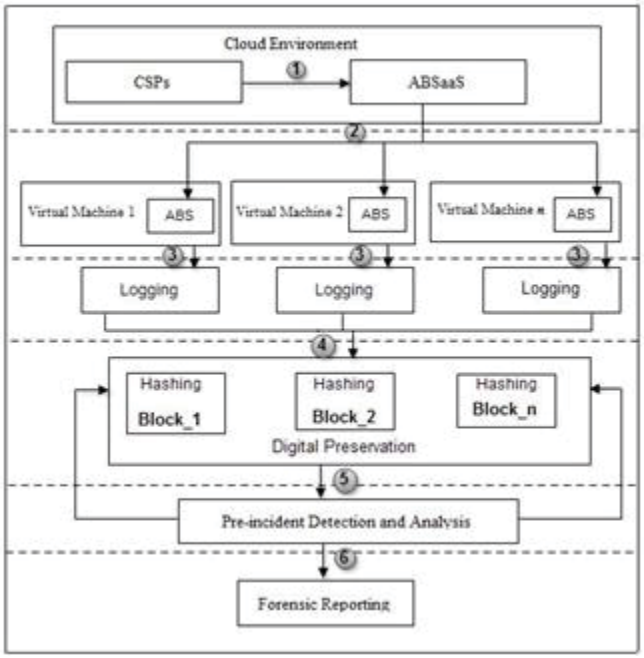


Figure 1: Highlight of distributed forensic client solution (Kebande & Venter,2015)

Figure 1 shows a block diagram of the NMB solution represented as a distributed forensic client. It consist of the cloud environment shown in the part labled 1, virtual machine *1,2..n* where the botnet as an agent based solution(ABS) is installed in part labeled 2. Thereafter, logging is done in step 3 where digital evidence is captured, hashed and digitally preserved in step 4. Step 5 allows detection of potential incidents from the

collected evidence in a readiness approach. Finally a forensic report is generated in step 6 that shows activities on the collected potential evidence.

3. Related work

Research work by Mouton and Venter (2011) proposed a prototype for achieving DFR on wireless sensor networks where demonstrations of a working prototype showed that DFR layer could be added to wireless network but the cloud was hardly a focus in this context. Next, a prototype for guidance and implementation of standardized digital forensic investigation process by Valjarevic and Venter (2014) enabled reliable logging of all actions within the process of the comprehensive and harmonized digital forensic investigation process model. In this prototype the authors focus was primarily on the Digital Forensic Investigation Process Model (DFIPM) and it was hardly focused in the cloud. A Reference Architecture for a Cloud Forensic Readiness System proposed a forensic readiness approach system that was able to implement Forensic readiness in the cloud, the benefit was to save time and money for digital investigation processes through the reduction of direct impact on cloud activities (De Marco, Ferrucci, & Kechadi, 2014). According to Dykstra & Sherman (2012) different forensic tools like EnCase have been analysed for Cloud Forensic Readiness (CFR) aspects. This analysis has shown that the data that can be collected by these tools might not be so reliable because of lack of important features that can help digital forensic investigators.

4. Prototype Implementation overview

In this section, the authors discuss how the concept of achieving DFR using the cloud NMB solution has been realized which has previously been described in Section 2. The authors first describe how digital evidence is extracted based on the processes including the CPU usage, RAM usage, IP addresses and keystrokes. Then the authors explain the components that have been implemented to support the cloud NMB solution towards achieving DFR. It is worth noting again that in this process the functionality of the existing cloud architecture is not tampered with because the collected PDE is manipulated outside the cloud environment. The need for the implementation of this prototype as mentioned in the introduction has been motivated by the rise of cyber-crime incidents, hence, the prototype serves as a proof of concept as a way of support for achieving DFR based on the ISO/IEC 27043: 2015 standard guidelines on forensic readiness

4.1 Technical goals

In order to find potential digital evidence that can link a suspect to a crime in the cloud environment, intruder's footprints are basically examined based on the system content and log files which are relatively manual forensic analysis. Furthermore, with the existence of big voluminous data, this proposition might be impossible when it comes to tracking the source of potential attacks because while filtering logs using manual forensics you might remove what you are looking for without noticing. In fact, since these processes are implemented separately it might be tedious to reconstruct the sequence of events in order to create a hypothesis that can be used in a court of law for digital investigations.

The above mentioned limitations can be overcome using the prototype that the authors have developed. The prototype focuses on the following technical goals: Monitors activities in the cloud environment using a distributed obfuscated NMB solution that acts as a distributed forensic client. Next, the digital information gathered in this process is digitally preserved then transmitted to a centralized forensic database for secure analysis. These processes are able to monitor the CPU usage in case a malware is utilizing the processing power that is allocated to each client computer, as well as the RAM processes and the keystrokes with respective timestamps. Lastly all events are reconstructed for easy detection of the causality of any potential security incident.

4.2 Digital forensic evidence collecting prototype

A Digital Forensic Evidence Collecting System (DFECP) is represented as an agent-based solution which acts as a botnet with modified functionalities that operates in a non-malicious fashion. DFECP is able to monitor activities by recognizing the CPU usage periodically, RAM usage, key-logging and processes dealing with data replication in the cloud. All this processes have been developed in order to help the cloud to be prepared forensically for digital investigations. It is worth noting again that these processes comply with forensic readiness processes that have been highlighted in the ISO/IEC 27043: 2015 standard. The DFECP will monitor data in motion and give reports periodically that show respective timestamps. The process is invoked by a forensic administrator and if

at any time the graphs seem unusual, all the collected activities that are associated to the users are linked to a respective IP address to enable reconstruction of events that can link a suspect to a given crime. For example, if a malicious operation is able to utilize the processing power, then the CPU graph might change tremendously and this information will be useful when a digital forensic investigation is triggered in a target system. The DFECP is able to collect volatile and non-volatile data which is digitally preserved in a forensic database as highlighted by Kebande and Venter (2015). The authors conducted the experiments using three client machines and one server, the client machines were running Windows 7 while the sever machine was running Windows 10 , the set-up is shown in Figure 2.The reports for the collected forensic evidence are generated by computer using the IP address or by the username. This has also been shown in Figure 2.

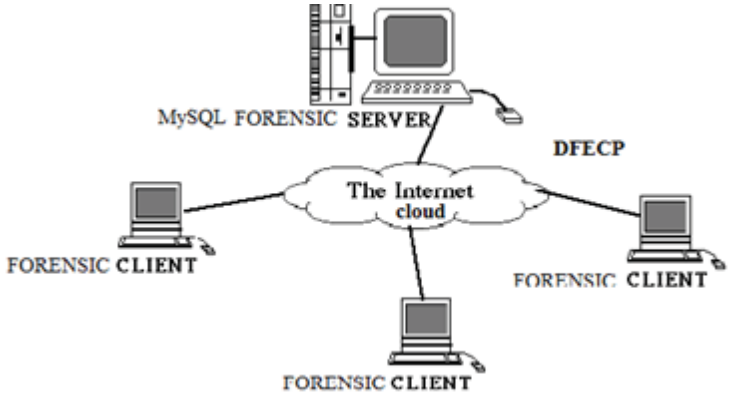


Figure 2: Architectural components of the prototype

Figure 2 shows the architectural components of the prototype which has been implemented to test the possibility of collecting potential evidence from the cloud environment by modifying the functionality of a botnet to act as a distributed forensic client. The functionality of the proposed prototype enables monitoring, digital preservation, pre-incident analysis, event reconstruction and forensic reporting. The prototype consists of MySQL forensic server that handles storage of the collected information. For easy maintenance and installation the DFECP functionality has been written using PHP, a server-side scripting language and C++ programming language. The main feature employed in DFECP is gathering digital forensic evidence using the distributed forensic client. The collected evidence is digitally preserved through hashing which is then stored in the database that with a sequence that shows the system User IP, processes, User Activities, Timestamps, CPU usage and RAM usage periodically. CPU and RAM usage can be used to monitor anomalous usage of processing power, memory, and huge amount of data that may be pushed to the cloud. After this process the collected data is hashed to preserve its integrity, finally events are reconstructed as highlighted by Kebande and Venter(2015c) through correlating evidence that may be ordered by system IP address, time or the process ID's. It is worth noting that the processes employed in this prototype comply with the forensic readiness processes that have been highlighted in the standard of ISO/IEC 27043: 2015.

Our envisaged approach has some drawbacks; firstly, implementing the prototype is marred by huge, complex and tremendous size of data that jets in large volumes. Even though the back-end storage is able to create a relational database to cater for this data, lack of enough storage might affect the provenance of the data object. Figure 4 shows the running processes after an NMB is deployed to act as forensic clients.

4.3 Prototype requirements

The requirements presented in this section have been used to portray the effectiveness of the proposed prototype. The prime objective of the prototype is to seamlessly enable the Non-Malicious Botnet to be deployed in the cloud environment to collect potential digital evidence for digital forensic readiness (DFR) purposes. A summary of the requirements have been shown in Table 1 coupled with an explanation.

4.4 Experimental results

Our approach focuses on how digital forensic evidence can be gathered from the cloud environment. The idea behind this experiment is to prepare the cloud forensically for digital investigations. Experimental results presented in this section show that DFR can be achieved in the cloud when a NMB is employed. This could only be achieved through the collection of forensic logs in a proactive process that could monitor the, CPU usage,

RAM usage, keystrokes and their respective timestamps. These experimental results presented reasonable performances because our implementation and design represented a practical mechanism.

Table 1: Summary of the requirements for achieving digital forensic readiness in the cloud (Kebande & Venter, 2016)

	Requirement	**Summary**
1	Forensic Logging capability and management	Forensic Logs to be used as digital evidence should be collected in a virtualised environment.
		It is important to know how logging is done, what is logged and when to log.
2	Integrity and Authenticity	The retained digital evidence should be digitally preserved.
		Verification authenticity should be possibe if there is a need for a digital forensic investigation.
3	Time stamping	Each log should have a time stamp to in order to prove its integrity.
		All events and activities should have time stamp representation.
4	Digital evidence characterization	Digital evidence should be grouped respective file format for possible incident identification.
		Activity analysis should be conducted to isolate potential security incidents.
5	Non-modification of existing cloud architecture	Functionalties of existing cloud architecture are not modified or tampered with.
		Activities like computation of evidence and analysis are conducted outside the cloud environment.
6	Security implementation	The software application solution should be protected other malicious activities.
		The software application should be deployed in a trusted environment
7	Obfuscation	Software aoolication's patterns are changed in a nonsensical manner to deter survellance.
		Obfuscation is enforced for privacy reasons
8	Event reconstruction	A hypothesis that should prove a fact in a court of law should be developed based on events.
		Structure and occurrence of events should be easily distingushed.
9	Legal requiements	The legal perspective and provisons across diverse jurisdictions should be known prior to a digital forensic investigation.
10	Forensic reporting	A readiness report that shows the interpretation process as a result of digital evidence retention should be generated.

On the one hand, Figure 3 shows a block of potential digital evidence that is captured when the NMB prototype is executed. The figure shows the total CPU and RAM usage as evidence is captured. After capture evidence is sent to the database using the POST/sendata.php HTTP/1.1 request that is shown in the Figure 3. Figure 4 on the other hand shows a the same block of log data stored in MySQL database. Also shown is the hash created as a mode of digitally preserving the log, the timestamp, IP address of the forensic client and Machine ID.

Figure 3: Block of captured potential digital evidence, CPU & RAM usage that is posted to the forensic database

Figure 4: Hash, timestamp, machine IP address, machine ID for stored potential evidence

The raw data that is shown in Figure 4 represents the block of captured potential digital evidence that has been posted to the forensic database. This can fully be seen inside a circle of Figure 5, that shows the entire log that can be used as potential evidence in a digital forensic readiness approach.

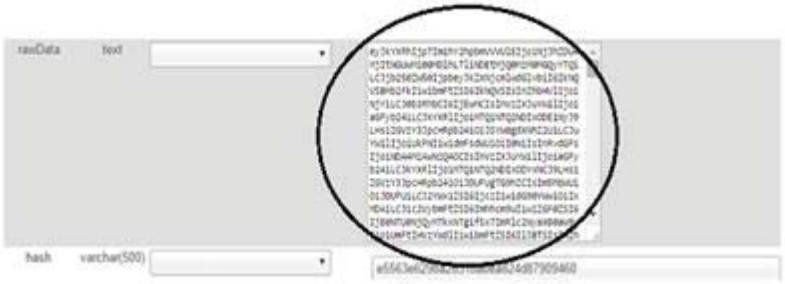


Figure 5: Sample potential evidence represented as rawData in the database

Figure 6 shows the keystrokes that are collected and posted to the forensic server by the NMB. Moreover, the collected value of the keystrokes is shown in the field labeled as value. Each id of the value pressed in the keyboard is also shown and the timestamp which is represented in milliseconds. Each of the value representing the keystroke is stored as digital forensic evidence. Notwithstanding that, Figure 7 shows performance results on the CPU usage after running the NMB solution on a set of clients in the cloud environment when the block of potential digital evidence is posted to the forensic database. The timestamp recorded from the report shows 2016-02-14 09.36.00, with IP address 196.248.115.230 which is the IP stored in the forensic database as shown previously in Figure 5.

id	name	username	value	total	description	date	logEntryId
10386	Keyboard	haron	k	0	Keystroke	1457004340436	87
10387	Keyboard	haron	l	0	Keystroke	1457004340653	87
10388	Keyboard	haron	k	0	Keystroke	1457004340659	87
10389	Keyboard	haron	j	0	Keystroke	1457004340731	87
10390	Keyboard	haron	d	0	Keystroke	1457004340743	87
10391	Keyboard	haron	s	0	Keystroke	1457004340753	87
10392	Keyboard	haron	c	0	Keystroke	1457004340821	87
10393	Keyboard	haron	l	0	Keystroke	1457004340849	87
10615	Keyboard	haron	[Enter]	0	Keystroke	1457004567780	90
10738	Keyboard	haron	w	0	Keystroke	1457004700695	93

Figure 6: Collected keystrokes stored in a forensic server

The report can either be generated using the computer or IP address or using the computer username as shown in Figure 8. For the sake of this paper, we have generated the report using the Computer, which displays the IP address. Nevertheless, potential digital evidence can be extracted based on the date. We used a start date of 2016-02-14 ant time 09:36:00 and an end date as 2016-02-14 12:54:00 and the result are shown using the CPU usage graph.

Figure 7: CPU utilization with time stamps graph collected by an NMB solution

RAM usage is also monitored as a block of digital data is posted to the forensic database. The RAM graph in Figure 7 has been generated based on the digital data that was previously pushed to the database as shown in Figure 3. The importance of the graphs that have been presented in Figure 7 and 8 are to monitor if there is any unusual activity that might consume the processor or memory while gathering digital evidence. Respective timestamps are shown in the graph of Figure 7 which shows CPU utilization, this are labeled *X, Y ,Z, V*. For example, the following parameters according to Jordan (2015) might create anomalies in the CPU energy consumption rate: CPU load, memory consumed, network packets received, network packets transmitted, disk reads and disk writes. Based on this premise Figure 8 shows the RAM usage graph report.

Figure 8: RAM utilization graph

The reader has been introduced to a practical forensic gathering prototype that is able to make the cloud forensically ready for digital forensic investigation as per the processes that have been define in ISO/IEC 27043: 2015 international standard. After collecting potential evidence, the authors monitored the CPU and RAM overheads that the prototype uses to gather digital evidence from various forensic clients and post it to the forensic database. Also monitored in this aspect is the effect of the tremendous amount of data that is taken into the forensic database after the POST request. In the next section, the reader is introduced to the legal and compliance aspects on admissibility of digital evidence.

5. Legal and compliance aspects on admissibility of digital evidence

Even though the aspect of digital evidence gathering is very important for digital forensic experts and law-enforcement purposes, different jurisdictions have different digital evidence perspectives with regard to admissibility. Around 92% of agencies have a legal authority to gather digital evidence, due to this some will require the national law permission for legal provision (DEG, 2010). For example, in the USA, Rule 702 of the Federal rules of evidence on the expert witness on admissibility is allowed to issue an opinion or a testimony at the trial based on the specialized knowledge he has, training or experience. If the evidence given is found to be reliable then the fact finder may reach a decision (Hutchinson, 2012). Discounting that, the Association of Chief of Police Officers (ACPO) of the UK, highlights in Section 4.6.3 that all the records for actions captured online should be kept as evidence (ACPO, 2009). This has a very big significance with digital forensic readiness. Additionally, the Electronic Communications and Transactions (ECT) Act 25 of 2002 of South Africa, Regulation

of Interception of Communications and Provision of Communication-related Information Act (RICA) 70 of 2002 of South Africa, Protection of Personal Information (POPI) Act 4 of 2013 of South Africa also highlights on provisions of digital evidence gathering. The laws may give a provision, a direct or indirect authority to relevant digital evidence that may be needed to conduct digital forensic investigations. A critical evaluation of the prototype is given in the next section.

6. Critical evaluation of the proposed prototype

The paper sought to provide digital forensic practitioners with techniques of gathering digital forensic information that can be useful during digital forensic investigation process through a Digital Forensic Evidence Collecting System (DFECP). The readiness processes from (ISO/IEC 27043: 2015) standard that have been employed in this paper serves as the easiest way of conducting digital forensic readiness in the cloud environment because the functionality of the existing cloud architecture is not modified at any instance, which saves on cost and time. The authors installed the prototype in three forensic client computers as targets that runs Windows 7. After analysis, we found out that the processes were running constantly and blocks of information was able to be collected periodically, until when we decided to push huge amount of data into the system. A report generated by the computer with IP address 196.248.115.230 and username john Doe with a start date 2016-02-14 and time 09:36:00 and an end date as 2016-02-14 12:54:00 shows collection of digital forensic evidence. CPU usage start at 60 from the graph shown in Figure 3, if then fluctuates up and down at the various intervals depending on what happens to the forensic client system.

The demonstrations done were able to show that DFR could be achieved by modifying the functionality of the botnet to act as a forensic-based agent solution that is capable of collecting forensic based evidence in the cloud. It is important to note again that the functionality of the existing cloud architecture is not modified because, activities and manipulation of the collected digital evidence as highlighted in a research paper by Kebande and Venter (2015b) is conducted outside the cloud. In such an instance it might warrant an investigator to isolate a cloud instance as mentioned by Delport, Khon and Olivier (2011).

The cost of conducting digital forensic investigation without employing DFR is higher. One would prefer to have a running application that is able to forensically prepare the cloud for post-event response after incident detection. As Rowlingson (2004) highlights, there is minimum disruption of business processes while conducting forensic investigation and based on this evidence that is related to potential digital crimes that has a possibility of affecting an organisation can be gathered. Having looked at various aspects on the prototype, the next section gives a conclusion and proposes a future work.

7. Conclusion and future

In this paper, the authors have presented a prototype that is able to gather digital forensic information in a proactive approach for DFR purposes. The prototype take advantage of being able to maximize the use of potential digital evidence while saving the cost and time of conducting digital forensic investigation. The system uses a modified botnet in the cloud environment that acts as an agent that is able to “infect” instances in the cloud and gathers potential digital evidence. Additionally, the authors use standardised forensic readiness processes that have been defined in the ISO/IEC 27043:2015 standard.

If we revisit the problem statement that the paper addressed, we see that it is possible to modify initially considered malicious software to conduct DFR process. Moreover, a software prototype can be used to conduct DFR process. Based on this aspect, our software prototype was able to collect digital information and store it for forensic readiness purposes.

The authors still plan to improve the prototype further however, future work that remains to be done include correlating data in a well formatted mode after capture, build a computation mechanism that is able to compute voluminous data across high speed networks. Nevertheless, the authors hope to standardize the process so that it can be widely accepted.

Acknowledgements

This work is based on research supported by the National Research Foundation(NRF) of South Africa (Grant-specific unique reference number UID85794). The Grant holder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF-supported research are

those of the author(s) and the NRF accepts no liability whatsoever in this regard. The authors wish to thank the ICSA Research Group of the Department of Computer Science at the University of Pretoria for the support towards coming up with this research.

References

Agrawal, D., Das, S., & El Abbadi, A. (2010). Big data and cloud computing: new wine or just new bottles?. *Proceedings of the VLDB Endowment*, *3*(1-2), 1647-1648.

ACPO - Association of Chief Police Officers (2009) Good Practice Guide for Computer Based Electronic Evidence.

"Apache Hadoop", http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F.-Accessed in January 2015

Beebe, N. L., & Clark, J. G. (2005). A hierarchical, objectives-based framework for the digital investigations process. *Digital Investigation*, *2*(2), 147-167.

Cohen, M. I., Bilby, D., & Caronni, G. (2011). Distributed forensics and incident response in the enterprise. *digital investigation*, *8*, S101-S110.

Cohen, F. B. (2010). Fundamentals of digital forensic evidence. In *Handbook of Information and Communication Security* (pp. 789-808). Springer Berlin Heidelberg.

Casey, E. (2009). Digital forensics: Coming of age. *Digital Investigation*, *6*(1), 1-2.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Delport, W., Köhn, M., & Olivier, M. S. (2011, August). Isolating a cloud instance for a digital forensic investigation. In *ISSA*.

Dittrich, J., & Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, *5*(12), 2014-2015.

De Marco, L., Ferrucci, F., and Kechadi, T. (2014) "Reference Architecture for a Cloud Forensic Readiness System", EAI Endorsed Transactions on Security and Safety, ICST, to appear.

DEG(2010) "Digital Evidence Gathering" accessed at "http://www.internationalcompetitionnetwork.org/uploads/library/doc627.pdf"

Draft NISTIR 8006 (2014) NIST Cloud Computing Forensic Science Challenges accessed at http://csrc.nist.gov/publications/drafts/nistir-8006/draft_nistir_8006.pdf

Gary Palmer (2001), "A Road Map for Digital Forensic Research". Technical Report DTR-T001-01, DFRWS, *Report From the First Digital Forensic Research Workshop* (DFRWS).

Grispos, G., Storer, T., & Glisson, W. (2012). Calm before the storm: The challenges of cloud computing in digital forensics. *International Journal of Digital Crime and Forensics*, *4*(2), 28-48.

Gartner(2015)"Special report on cloud computing." Accessed at http://www.gartner.com/technology/research/cloud-computing/report/

Hutchinson, C. T. (2012). What Is an Expert Witness?.

ISO/IEC 27043, "Investigation Principles and Processes", unpublished draft international standard (2012).

International Data Corporation (IDC) Top 10 prediction 2014. Accessed at "http://www.idc.com/research/Predictions14/index.jsp;jsessionid=0F5F27EF62AF965596E9D2177A95391B"

Kebande, V. R & Venter, H. S. (2014a). A Cloud Forensic Readiness Model Using a Botnet as a Service. In *The International Conference on Digital Security and Forensics (DigitalSec2014)* (pp. 23-32). The Society of Digital Information and Wireless Communication.

Kebande, V.R & Venter, H.S (2015a, February). Obfuscating a Cloud-Based Botnet Towards Digital Forensic Readiness. In *Iccws 2015-The Proceedings of the 10th International Conference on Cyber Warfare and Security* (p. 434). Academic Conferences Limited.

Kebande, V. R., & Venter, H. S. (2014b). A Cognitive Approach for Botnet Detection Using Artificial Immune System in the Cloud.

Kebande, V., & Venter, H. (2015d, October). Towards a Model for Characterizing Potential Digital Evidence in the Cloud Environment during Digital Forensic Readiness Process. In *ICCSM2015-3rd International Conference on Cloud Security and Management: ICCSM2015* (p. 151). Academic Conferences and publishing limited.

Kebande, V. R., & Venter, H. S. (2015c, August). Adding event reconstruction to a Cloud Forensic Readiness model. In *Information Security for South Africa (ISSA), 2015* (pp. 1-9). IEEE.

Kebande, V., & Venter, H. S. (2015b, July). A Functional Architecture for Cloud Forensic Readiness Large-scale Potential Digital Evidence Analysis. In *Proceedings of the 14th European Conference on Cyber Warfare and Security 2015: ECCWS 2015* (p. 373). Academic Conferences Limited.

Kebande, V., & Venter, H. S. (2016, March). Requirements for Achieving Digital Forensic Readiness in the Cloud Environment using an NMB Solution. In *Proceedings of the 11th International Conference on Cyber Warfare and Security 2015: ICCWS 2015*. Academic Conferences Limited-To appear.

Kohn, M. D., Eloff, M. M., & Eloff, J. H. (2013). Integrated digital forensic process model. *Computers & Security*, *38*, 103-115.

Mouton, F., & Venter, H. S. (2011, September). A prototype for achieving digital forensic readiness on wireless sensor networks. In *AFRICON, 2011* (pp. 1-6). IEEE.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

Reddy, K., & Venter, H. S. (2013). The architecture of a digital forensic readiness management system. *Computers & Security*, *32*, 73-89.

Rowlingson, R. (2004). A ten-step process for forensic readiness. *International Journal of Digital Evidence*, *2*(3), 1-28.

Shropshire, J. (2015, October). Securing Cloud Infrastructure: Unobtrusive Techniques for Detecting Hypervisor Compromise. In *ICCSM2015-3rd International Conference on Cloud Security and Management: ICCSM2015* (p. 86). Academic Conferences and publishing limited.

Shafer, J., Rixner, S., & Cox, A. L. (2010, March). The Hadoop distributed filesystem: Balancing portability and performance. In *Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on* (pp. 122-133). IEEE.

Tan, J. (2001), "Forensic readiness", Technical. Cambridge USA: @stake, Inc.

Yasinsac, A., & Manzano, Y. (2001, June). Policies to enhance computer and network forensics. In *Proceedings of the 2001 IEEE workshop on information assurance and security* (pp. 289-295).

Van Staden, F., & Venter, H. (2012). Implementing Forensic Readiness Using Performance Monitoring Tools. In *Advances in Digital Forensics VIII* (pp. 261-270). Springer Berlin Heidelberg.

Wen, Y., Man, X., Le, K., & Shi, W. (2013, May). Forensics-as-a-Service (FaaS): Computer Forensic Workflow Management and Processing Using Cloud. In *CLOUD COMPUTING 2013, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization* (pp. 208-214).

Valjarevic, A., Venter, H. S., & Ingles, M. (2014, August). Towards a prototype for guidance and implementation of a standardized digital forensic investigation process. In *Information Security for South Africa (ISSA), 2014*(pp. 1-8). IEEE.

Dykstra, J., Sherman, A.T. (2012) Acquiring Forensic Evidence from Infrastructure-as-a-Service Cloud Computing: Exploring and Evaluating Tools, Trust, and Techniques. In Proceedings of the 12th Annual DF Research Conference (DFRWS'12), Washington, DC, USA (Digital Investigation), August 2-8, vol. 9, pp. 90–98

Stability of Iris Patterns in Different Parts of the Visible Spectrum

Ľuboš Omelina[1,2,3], Bart Jansen[1,3], Alexandra Biľanská[2] and Miloš Oravec[2]
[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, Belgium
[2]Institute of Computer Science and Mathematics, Slovak University of Technology in Bratislava, Slovakia
[3]iMinds, Dept. of Future Health, Belgium
lomelina@etro.vub.ac.be
bjansen@etro.vub.ac.be
alexandra.bilanska@gmail.com
milos.oravec@stuba.sk

Abstract: Biometric recognition becomes an integral part of devices that we have in daily use. Due to its increased comfort and reliability, fingerprint recognition is now part of the high-end personal devices. Iris recognition offers superior accuracy, increased comfort in enrolment, and in case of colour cameras could be used, it would work without additional hardware. With additional hardware (NIR cameras and a NIR light source), it is possible to improve the standard iris recognition method and to achieve higher recognition accuracy. However, using visible spectrum in combination with up-to-now embedded colour cameras, could transform millions of existing devices to iris-enabled scanners and take advantage of increasing trends in mobile biometrics. We study the possibility of using cameras already embedded in mobile phones for the iris recognition task, particularly we focus on the feature extraction phase with state-of-the-art hardware. In this paper we analyse iris patterns extracted from colour images. Different parts of the visible spectrum are used, in order to evaluate the most suitable parameters in the iris recognition task. Due to the lack of public databases, we collected a database of iris images from 20 people with consumer mobile phone cameras. We explore different parameters of the feature extraction method proposed by John Daugman. The experimental results show the potential of multimodal combination of different features per spectrum.

Keywords: iris recognition, mobile, visible light

1. Introduction

A common objection to using fingerprint security for mobile devices was that a fingerprint could be stolen and misused. Users touch devices with their fingers, leaving an easy way to reconstruct them by an attacker. In contrast, using the iris for recognition leaves no trace marks on a device and offers superior recognition accuracy. Recent research suggests that the iris can be accurately recognized even from colour images (Boyce et al 2006, Raja et al. 2015) and that raises the question whether cameras already embedded in the mobile devices could be used to recognize people based on their irises. In this paper we analyse the potential of performing iris recognition from images collected with mobile devices and report the accuracy using state-of-the-art iris recognition methods. We use probably the most successful iris recognition method proposed by J. Daugman (2004). Daugman's method relies on near infra-red images (NIR), due to the stability of the iris pattern under this modality. In contrast to visible light (VIS), irises exhibit very high reflectance in the NIR channel and decrease with the wavelength of the light (mainly the brown coloured irises). In (Boyce et al, 2006), the authors studied the role of information contained in individual spectral channels (also) of VIS. Part of their work was also done by M. Monaco (2007) where he presented the potential of using VIS images for iris recognition.

Most public datasets containing colour images (also in (Boyce et al,2006)) were collected by specialized cameras and within a laboratory constrained environment. In an unconstrained environment however, the VIS images typically contain significant amounts of specular reflections, might not be well lit and contain higher amounts of noise. Few databases contain images captured from a variety of regular smartphones in unconstrained environments (e.g. (De Marsico M. et al 2015)) and do not focus on images taken at very short distances from the camera. Our goal was to create a challenging database that could reveal limitations of iris recognition from mobile handheld devices. We collected a database of iris images from 20 people with consumer mobile phone cameras. Since the NIR light is not within the VIS spectrum we studied the impact of using different parts of the visible spectrum on the accuracy of the iris recognition task.

2. Methods

Iris recognition methods and systems rely on NIR images due to the stability of the iris pattern under this imaging modality. By using VIS images, the recognition becomes a much more challenging task mainly due to specular

reflexions and variability of lighting conditions. Iris codes extracted from visible images have thus higher volatility and provide lower recognition rates when standard iris recognition algorithms are used. In this section we describe properties of sensors typically used in mobile devices and their relation to multispectral imaging within the VIS spectrum.

2.1 Iris recognition

The iris recognition process follows the general biometric process introduced by A. K Jain et al (2004). The process begins with an image acquisition and its preprocessing. In our case, we manually localize the iris boundaries and eyelids. Then the image is segmented and the iris texture is transformed from Cartesian coordinates to the polar coordinate system (see the unwrapped iris in Figure 1). From this unwrapped representation, features are extracted using Gabor filters and the result is encoded to a binary template. For matching two different irises the Hamming distance is used. This method proposed by J. Daugman (2004) is de-facto a standard and is applied in vast majority of iris recognition systems today. More details on state of the art methods in iris recognition was described by Burge and Bowyer (2013).

a)  b) 

Figure 1: Example of the iris image. a) Original image from the captured database, b) Unwrapped iris in the polar coordinate system

2.2 Mobile devices and image acquisition

A VIS colour image contains information across multiple bands of the electromagnetic spectrum (400-700 nm). Specific bands are captured by different colour filters into different pixels. Most of the consumer cameras use 3 different filters into a Bayer pattern (see Figure 2). In addition, sensors include an NIR blocking filter to prevent NIR band having any influence on the image. There are twice as many green filters as red or blue to adjust images to the perception of the human eye. The devices contains an IMX214 sensor which acquires the spectral information as follows: (a) blue channel with peak at 470nm, (b) green channel with peak at 520nm, (c) red channel at 600 nm (see Figure 3). In this paper we will refer to individual spectral channels in terms of colour models. The colour model is defined as a composition of three values related to different parts of VIS spectra. Thus these three values define a frequency range and an intensity of a colour depending on the used sensor.

Compared to the static mounted iris scanners, cell phones are significantly lighter - hand held image capturing on such devices results into a greater amount of camera movement. Another difference is that mobile devices usually use a CMOS sensor with rolling shutter, but the iris scanners use CCD sensors with a global shutter. In case of rolling shutter, the scene is not captured at once, but rather scanned per row or column. Camera shake together with rolling shutter have severe negative impact on the captured iris.

For capturing the irises we used the standard camera applications to capture the full resolution images with automatic settings. It is important to note that the automatic settings may not be ideal for iris capture.

For our study we collected a database in an unconstrained environment with 5 different smartphones. For collection of the samples we used automatic settings of the camera – autofocus, auto-exposure and auto-white balance (aka 3A algorithm). The iris regions in the images were manually labelled in order to avoid errors introduced in segmentation process. Manual segmentation (Figure 1a**Error! Reference source not found.**) included labelling of inner and outer boundaries of the iris (modelled as ellipses) and position of eyelids (modelled as parabolas).

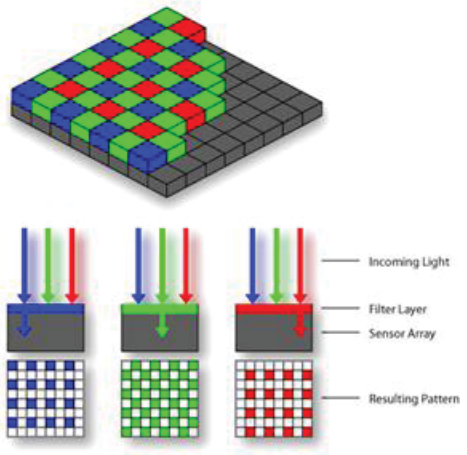


Figure 2: The Bayer arrangement of colour filters on a sensor[1].

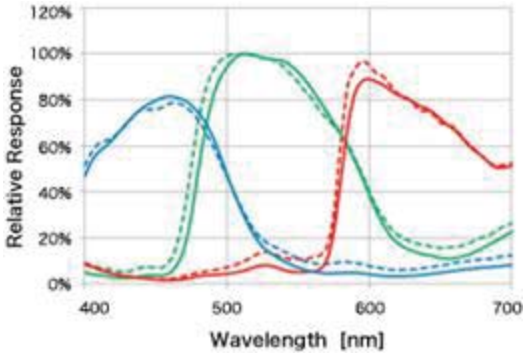


Figure 3: Spectral sensitivity characteristics of the Sony IMX214 (Solid line) and IMX135 (dashed line) sensors.[2]

3. Experiments

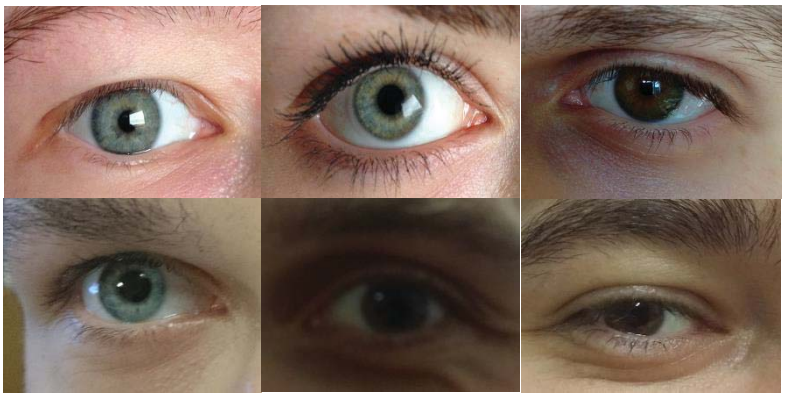


Figure 4: Examples of images contained in the collected database. The top row – correctly recognized images. The bottom row - images that could not be recognized correctly (due to noise, out of focus or too dark iris structure).

In our experiments we used 1 training sample per class (40 in total) meaning that the training samples were stored in the database as templates. We analysed each part of each colour model separately in a closed set experiment where we matched the nearest neighbour of the iris templates in the training set (360 samples). For each colour component of each colour model we performed grid optimisation of the Gabor filter scale (size measured in pixel). Orientation and scale were kept constant in our experiments. In our results we report the

[1] Source of the image: Wikipedia, The Free Encyclopedia, s.v. "Bayer filter", (accessed February 3, 2016), https://en.wikipedia.org/wiki/Bayer_filter

[2] Source of the image: Sony product information on IMX214, (accessed February 3, 2016), http://www.sony.net/Products/SC-HP/new_pro/april_2014/imx214_e.html

accuracy for only for the best size of the Gabor filters. The overall recognition accuracy together with the size of the Gabor filter for each colour model is shown in Figure 5. Results show that recognition with the YIQ colour model and the red component of the RGB colour model (R-RGB) achieved the highest accuracy of 80.37%. Detailed analysis shows that most of the irises that were not recognized correctly exhibit very poor quality (see the bottom line of Figure 4 for some examples). An interesting observation is that authentication based on any of the three components of the RGB colour model reached similar accuracy i.e. 80,37 % (the red channel), 78,48 % (the green channel) and 76,58 % (the blue channel). The best recognition accuracy for each of the three R, G and B components we measured for different (increasing) size of Gabor filters i.e. 6, 10 and 15 pixels. This suggests that separate channels of irises in RGB contain different features suitable for recognition hence are suitable for multimodal iris recognition. We also explored recognition rates when encoding irises with each complex part Gabor filter separately (see Figure 6). In contrast to RGB, in case of YUV only the Y/brightness component seems to be suitable for recognition. The two chrominance components (U and V) exhibit only poor results.

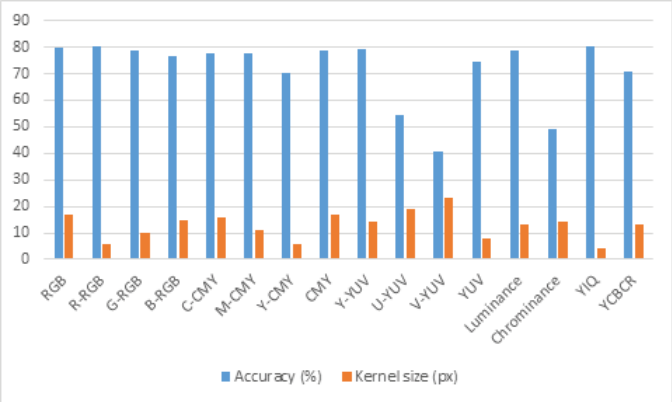


Figure 5: Recognition accuracy for different components of different colour models together with the best Gabor filter scales for component

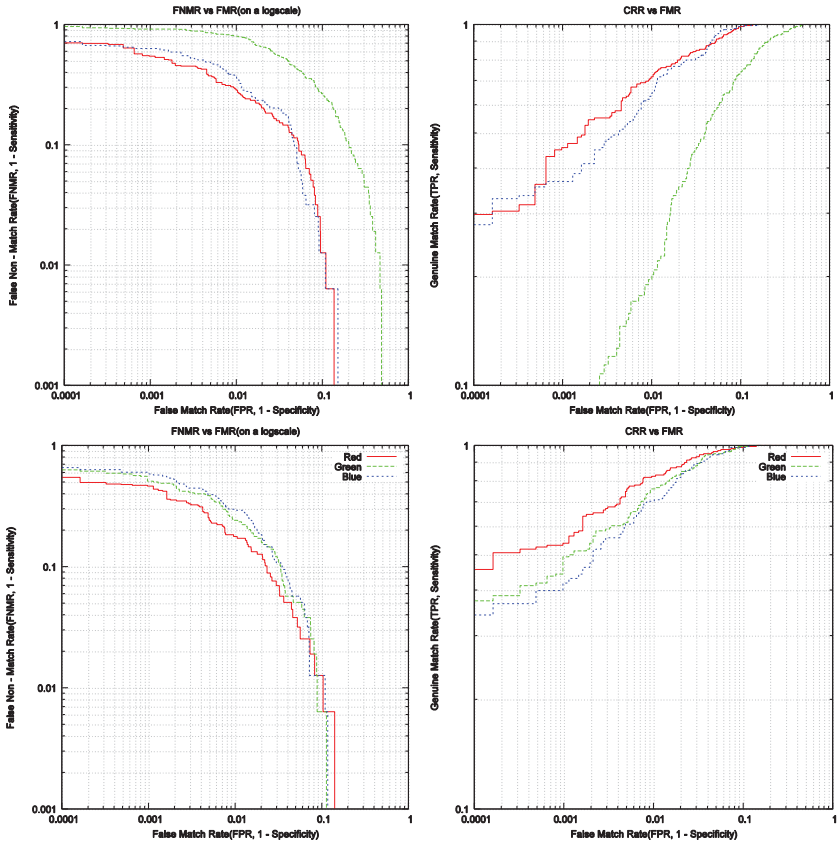


Figure 6: ROC curves indicating the recognition performance for different components of the RGB colour model for the real (top) or imaginary (bottom) part of the Gabor kernel

4. Conclusion

In this paper we analysed the potential of performing iris recognition from colour images taken by mobile devices. For this task we collected a database containing 400 images of 40 classes and 20 people using 5 different mobile devices. We conclude that spectral frequencies closer to NIR light are more suitable for iris recognition. Intensity of the light reflected from the iris (the luminance, the Y component from YUV colour model) has higher importance since it could outperform the colour information (the chrominance) by almost 30%. Results show that even in an unconstrained environment it is possible to recognize the irises with 80.4 % recognition accuracy. Our experiments also suggest that the optimal size of the Gabor filter used for extraction of the iris codes, depends on the used colour spectrum. An important limitation of our study is the automatic settings of the camera parameters that might not have been optimal, however we focused on the suitability of different spectral parts of VIS. The optimal settings (exposure, sensitivity) will be the subject of our future work. Future work may focus also on the specular reflections that are still an open problem in the VIS iris images taken in unconstrained environment.

References

Boyce Ch., Ross A., Monaco M., Hornak L., Li X.(2006) *Multispectral Iris Analysis: A Preliminary Study* in Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on, pp.51-51, 17-22

Daugman, J.(2004) *How Iris Recognition Works.* IEEE Transactions on Circuits and Systems for Video Technology. 14,21–30.

Jain A.K., Ross A., Prabhakar S.(2004) *An Introduction to Biometric Recognition.* IEEE Transactions on Circuits and Systems for Video Technology.14,4–20.

Monaco M.K. (2007) *Color Space Analysis for Iris Recognition*, MSEE Dissertation, West Virginia University

Bowyer K.W., Hollingsworth K., Flynn P.J.(2008) *Image understanding for iris biometrics: A survey.* Computer Vision and Image Understanding. 110,281–307.

Raja, K. B., Raghavendra R., Vemuri V.K., Busch Ch.(2015) *Smartphone based visible iris recognition using deep sparse filtering*, Pattern Recognition Letters,Volume 57,Pages 33-42

Marsico M., Nappi M., Riccio D., Wechsler H.(2015) *Mobile Iris Challenge Evaluation (MICHE)-I*, biometric iris dataset and protocols, Pattern Recognition Letters,Volume 57,Pages 17-23

Grid Security Policy Monitoring System (GridSPMS): Towards Monitoring the Security Dimension of Grids

Abdulghani Suwan and Francois Siewe
School of Computer Science and Informatics, Faculty of Technology, De Montfort University, UK
p02318242@dmu.ac.uk
FSiewe@dmu.ac.uk

Abstract: Grid computing systems are complex and dynamic environments requiring appropriate automated management mechanisms, which enable stable and reliable operation of the whole grid ecosystem. Moreover, the recent novel concept of mobile grid computing – a combination of grid systems and mobile devices – also requires new ways of monitoring the emerging security concerns associated with the ever-increasing role of network connections in mobile grids. Existing grid monitoring systems, albeit suitable for traditional, localised grid systems, seem to ignore the security dimension and do not offer appropriate support for enforcing security policies within a distributed grid system enhanced with mobile devices. Accordingly, this paper presents the Grid Security Policy Monitoring System (GridSPMS), a novel grid monitoring framework that extends traditional support for data monitoring, for instance tools and protocols, with mechanisms for collecting run-time data within grids, focusing on security. By doing so, GridSPMS aims to monitor the activity within a grid system and detect situations in which security policies have potentially been violated and, therefore, appropriate response actions must be taken. In this way, GridSPMS has the potential to counter security threats in the domain of mobile grid computing, and provide grid administrators with support for timely detection of and response to security-related incidents.

Keywords: grid computing, security policy, monitoring system

1. Introduction

With the emergence and rapid development of cloud computing in the last decade, grid computing is now primarily utilised in the context of academic research, receiving little, or no, interest or support from the IT industry. Traditionally, grids have not been considered a commercial product to be offered to a wide range of customers (Foster et al, 2008). Unlike cloud computing, which is a widely known, highly commercialised and industry-driven research area, grid computing is attracting less and less attention from the research community.

This situation, however, might change following the recent introduction of mobile grids. Mobile grid computing is an emerging computing paradigm positioned at the intersection of two research areas, namely, grid computing and mobile computing (Litke et al, 2004). It can thus be seen as a reincarnation of conventional grid systems, and its primary function is to extend the traditional capabilities of grids - that is, the provision of a large pool of aggregated computational and storage resources, in order to address computationally-intensive tasks (Foster and Kesselman, 1999) – with the computational capabilities of mobile devices provided via the network. Broadly speaking, a mobile grid can be defined as a complex scalable distributed system involving a variety of nodes, some of which might be mobile, geographically distributed around the globe, and use various communication protocols. From this perspective, mobile grid computing can be seen as an evolution of the concept of grids, from traditional, on-premises deployments to a distributed computer architecture, consisting of both server clusters residing in a data centre and multiple mobile devices, such as smartphones, tablets and laptops, connected to the main cluster via a wireless network.

With regard to computing technology in general, what was regarded as impractical or impossible just a few years ago, today is attracting increasing attention and investigation both from industry and academia. For example, portable mobile devices are becoming more and more powerful in terms of their CPU and memory capabilities, and the bandwidth of 5G mobile networks is approaching 1GB per second. Thus, the potential integration of grids within mobile computing is even more promising. The inherent benefit of mobile grid computing is that it allows users to access grid resources from a mobile device virtually anywhere, at any time, without needing to be near a hub. Moreover, in mobile grid architectures, mobile devices are expected to act not just as resource consumers, accessing grid resources, but also as resource providers in their own right, contributing to the shared pool of available resources (Katsaros and Polyzos, 2008). This latter feature may enable the formation of 'ad-hoc' mobile grid compositions on the spot. This makes it possible to create local, short-term mobile grids at various venues and locations where there is an increased concentration of smart mobile devices, such as conferences and universities (Litke et al, 2004).

Along with the aforementioned promising opportunities for increased computational capabilities and ubiquitous network access to computational and storage resources, come new emerging challenges, for example how to manage the resulting complex environments and maintain a stable quality of service (QoS). The issue of delivering the promised QoS to mobile grid users is a multi-faceted one, and can be considered from several perspectives. In the first instance, QoS depends on the networking capabilities of a mobile grid system. Mobile network bandwidth remains the primary factor in guaranteeing smooth and undisrupted data exchange between mobile devices and grid servers. Another important concern when it comes to satisfying QoS requirements is the implementation of the resource management mechanism responsible for resource discovery and selection, job scheduling and replication, data migration, and monitoring (Bichhawat and Joshi, 2010). In the presence of an increasing number of computational nodes constituting the mobile grid system, it is important to organise how various computational tasks are split and scheduled properly, so that they are executed in an efficient and scalable manner.

Finally, the emergence of mobile grid computing also poses new challenges concerning how these distributed systems should be properly protected and secured. As opposed to traditional grid architectures, where networking is not regarded as an important issue to take into account, in mobile grids the networking dimension plays a dominant role. The wireless nature of network connections introduces new security risks and means the resulting mobile grid systems are vulnerable to a wider range of threats, such as eavesdropping, data tampering and data tracing (Bichhawat and Joshi, 2010). Accordingly, appropriate novel monitoring and detection mechanisms are required to address these security issues and equip mobile grid systems with sufficient self-protective capabilities to maintain the required QoS, and at the same time provide sufficient scalability, which is a key feature of mobile grids.

In light of these considerations, this paper will argue that a potential method to support security management in mobile grid computing is to monitor grid activities, with the goal of detecting and reporting security-related incidents, followed by the generation of an effective security incident response plan. Accordingly, it will also present and explain the *Grid Security Policy Monitoring System* (GridSPMS), which is an original framework that is proposed by the author to monitor security policy compliance in grids, detect potential violations, and report on any detected incidents. The paper will describe the high-level architecture of the framework, and discuss its potential benefits and limitations.

The remainder of the paper will be organised as follows. Section II will introduce and briefly outline the security issues associated with grid computing in general, and in mobile grids in particular. This section will also provide an up-to-date overview of the research area and identify limitations and gaps. Then, to address the identified gaps, Section III will present the proposed GridSPMS and describe its conceptual architecture in a top-down manner, whereby a description of the conceptual architecture will be followed by a more detailed description of the individual components. Section IV will contain some concluding remarks and summarise the paper.

2. Motivation: Insufficient security in mobile grids

Arguably, security, albeit a major research challenge in IT in general, has never been a primary concern with regard to grids. Historically, grids have been associated with relatively small-scale, localised deployments within a single data centre (Bessis et al*,* 2010), where the number of network connections is limited and stable. Also, they have primarily served scientific purposes, and therefore the number of users accessing grid resources has typically been limited, and so could be easily controlled; as such, there has hitherto been little need or demand for enforcing various security checks. Moreover, grids have not been seen as a commercial product to be offered to a wide range of customers (Foster et al, 2008), and, consequently, there has been relatively little interest and support from the IT industry. As a result, the development of grids, including investigation of security-related issues, has been primarily driven by academic researchers (Foster et al, 2008).

For example, Foster et al (1998) were among the first to identify and address the challenge of ensuring the security of grid systems. The authors proposed and implemented an architecture for secure grids, and established theoretical and practical foundations for designing and implementing grid environments in a secure, robust and reliable manner. The architecture they proposed now serves as a reference model for other approaches aiming to achieve high levels of security in grids.

However, implementing grid systems in a secure manner, following Foster et al's (1998) proposed architecture on its own is not enough. An inherent requirement for achieving sufficient levels of security in grids, and in computing systems in general, is to implement appropriate monitoring mechanisms that enable the prompt and timely detection of potential violations. These monitoring mechanisms would collect various security-related data within grids, analyse that data in reference to a set of security policies, and determine whether there are any violations requiring immediate responsive action from the human administrator. Taken together, these processes are referred to as 'security policy monitoring' (Foster et al, 1998).

This latter requirement to implement security policy monitoring mechanisms, however, is beyond the general capabilities of existing grid monitoring systems. As indicated by a recent survey carried out by the authors (Suwan et al, 2016), existing grid monitoring frameworks appear to neglect the security dimension. As such, their capabilities with regard to integrating security-related monitoring metrics and policies are similarly limited.

With the rapid development of mobile technologies and the emergence of mobile grid computing, this is becoming an increasingly urgent challenge, which can be seen as being part of a global paradigm shift in the IT industry. Consumers are moving away from personal desktop computers toward portable mobile computing, supported by the powers of 'the cloud'. In 2011, smartphone sales exceeded PC sales for the first time (Cocotas, 2012), and Apple CEO Tim Cook recently declared that the PC is now "officially dead" (Ghosal, 2015). As the mobile revolution progresses, the issue of mobile security has been attracting much attention, and the market for portable and mobile devices would not be demonstrating exponential growth if it was not adequately supported by advances in security and data protection. Accordingly, as the mobile technology has advanced and the number of mobile devices has grown, the research and industrial communities have responded by putting more and more effort into investigating possible approaches to providing efficient and secure solutions for mobile platforms. In the 1990s, mobile technology was in its infancy and security issues were minor concerns to which little or no attention was paid . Then, in the 2000s, the community became focused on enabling mobile platforms with reactive self-protective mechanisms (Dunnewijk and Hultén, 2007) that would counter the potentially harmful consequences of detected malware. Now, more and more proactive mechanisms are attempting to predict and prevent possible attacks and threats before they happen (Finneran, 2013).

These advances in mobile security, however, have not yet been applied in the domain of mobile grid computing. As such, the next section of this paper will present the Grid Security Policy Monitoring System (GridSPMS), the authors' attempt to create a grid monitoring framework that fully supports the security dimension of grid computing.

3. GridSPMS: Conceptual architecture

In order to address the identified gaps in the domain of mobile grid security, the authors are currently developing a novel Grid Security Policy Monitoring System (GridSPMS). The system is built into an already existing grid monitoring framework, Ganglia[1], and extends its functionality to support the security dimension of grid computing. Before describing the actual design and architecture of GridSPMS, this section will first present a non-exhaustive list of desirable features of this kind of framework, which were taken into consideration and influenced the design of the final system.

These important features, which relate to both functional and non-functional properties, are:

- *Support for monitoring and enforcing security policies* is a fundamental feature of the envisaged monitoring framework. In a broad sense, a security policy is a set of rules that define and enforce certain access constraints on grid resources. For instance, security policies can be applied to limit access to memory and CPU resources per user, and restrict the creation of excessive jobs, and so on. Typically, security policies can target either individual users, or whole user groups, for example user roles.
- *Support for security incident response policies* is another key feature of the proposed grid monitoring framework. Security incident response policies describe reactive actions that should be taken whenever a security policy is violated – that is, whenever there is a security threat or incident. In the context of grids, an example of a security incident might be unauthorised access to computational resources, excessive consumption of resources, accessing resources using an insecure network connection, and so on.

[1] Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and grids (http://ganglia.sourceforge.net/).

Accordingly, security incident response policies may define reactive actions as ranging from simple reporting to the human administrator to more sophisticated actions, such as automatically initiating appropriate actions to return the system back to its initial stable state. The former is typically referred to as a *passive* response, and the latter is known as an *active* response.

- *Ability to seamlessly integrate with Ganglia* refers to the intention to adapt an already existing and efficient grid monitoring framework, as opposed to re-inventing the wheel. To a certain extent, the intention is to develop a fully compatible extension to the core Ganglia functionality, which will benefit from data collected and supplied by Ganglia to offer additional security-related capabilities.
- *Ease of deployment* is a requirement closely related to and stemming from the previous stipulation. By inheriting Ganglia functionality, the proposed monitoring system is expected to be easily deployable and configurable within grid environments.
- *Information abstraction* means that the monitoring system is able to integrate, synthesise and interpret collected information so that only relevant data is displayed to the user. Additionally, users are not supposed to interact with the back-end mechanism, which will be transparently abstracted by the front-end interface, namely the web browser.
- *Dynamic information retrieval and near real-time operation* are two features that enable the system to function in a timely manner. Data collected by the GridSPMS is highly dynamic in nature, and the system must be properly equipped to collect and process this data with minimum delay. This in turn will facilitate the prompt detection of potentially critical situations, and timely reporting of security incidents. In such data-intensive environments as (mobile) grids, where observed values may become obsolete and outdated within seconds, these requirements are seen by interested parties (i.e., grid providers and consumers) as one of the primary challenges (Sagiroglu and Sinanc, 2013).
- *Uniform data representation* is expected to overcome existing heterogeneity in data representation formats. There must be a common vocabulary of terms, that is, a standard way of representing information in the context of managed grid environments, which may include, for example, various grid components, system users, geo-locations, historical values, security incidents, and so on.
- *Support for scalability* is an inherent requirement of the grid domain, which is characterised by a highly distributed environment, consisting of a large number of computational nodes and clusters. With the introduction of mobile grids, this challenge is taken to another level, where data must be collected from multiple remote mobile nodes over the network. The envisaged monitoring framework is expected to support the integration of newly added nodes in an automated manner, so that changes are immediately reflected in the system in a transparent and seamless way.
- *High performance and robustness* are pre-requisites to enable scalability of GridSPMS, which should take into account the demands of a distributed environment to enable failure-free execution. This also implies that grid elements, for example individual nodes or whole clusters that are currently unavailable, have crashed or are under-performing should be safely isolated from the 'healthy' grid in order to support the stability of the whole grid ecosystem.
- *Ubiquitous remote access over the Internet* is another inherent requirement of mobile grids. With the integration of portable mobile devices into the grids, it becomes essential to enable remote access to the presentation layer of the monitoring system via some kind of 'mission control'. In other words, a user will expect to be able to remotely access and view the grid monitoring dashboard from any remotely-located device, at any time. Aside from viewing the current status of the monitored grid system, the user should also be able to manage and control the system.

Taken together, these features have been thoroughly considered when devising the conceptual architecture of GridSPMS. The design has also been influenced by the established practices of engineering complex monitoring systems in a modular, distributed and autonomic manner. More specifically, the proposed approach was inspired by surveying existing grid monitoring systems (Suwan et al, 2016) – their potential benefits and shortcomings were taken into consideration and underpinned the design of GridSPMS.

3.1 GridSPMS and the managed grid environment: four-layered conceptual architecture

This section will describe the actual architecture of the proposed monitoring system, in a top-down manner. First, it will present a high-level conceptual overview of GridSPMS and the managed grid environment, which

consists of four layers, and then go into more detail by explaining the individual components constituting the system.

Figure 1 schematically depicts the proposed four-layer architecture of the grid environment, managed by GridSPMS. These conceptual layers are interdependent, so that upper layers build and depend upon the lower layers. This organisation reflects the established structure of modern grid systems, in which low-level monitoring data is first collected from individual nodes, and then proceeds through a series of abstractions to represent an aggregated and synthesised view of a cluster of nodes, and, finally, of the whole grid. This approach represents the grid as a hierarchical network of clusters and nodes, and facilitates scalability. The information flow can be represented by the following triple:

Node → Cluster → Grid

The triple pattern also helps to uniquely identify individual nodes within GridSPMS. For example, to locate a specific node within a grid, it is necessary to locate its cluster first. From this perspective, grids and clusters act as namespaces – that is, nodes with same identifiers are allowed, provided they belong to different clusters and grids.

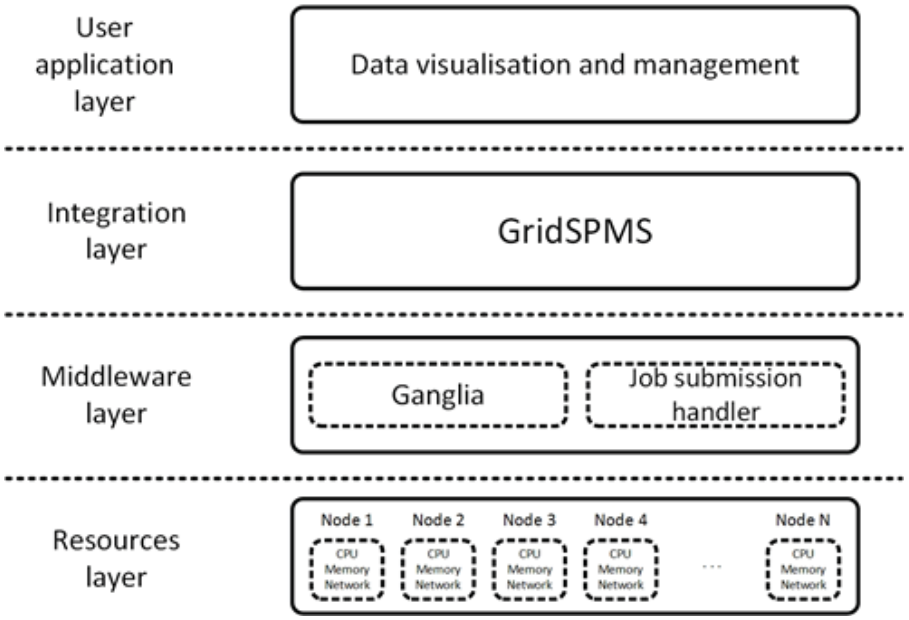


Figure 1: Four-layered architecture of the GridSPMS monitored environment.

- *The resources layer* represents monitored resources, such as individual nodes, clusters, network connections, and storage devices, equipped with a Ganglia agent to collect relevant data. Monitored metrics typically include the general availability of a particular resource, as well as CPU and memory utilisation, network statistics, occupied storage space, and so on. Essentially, this layer represents the application scope of the proposed GridSPMS – that is, traditional infrastructure resources belonging to this layer are the main subject for performing access control and security monitoring by GridSPMS.
- *The middleware layer* represents the Ganglia framework, which is responsible for a wide range of activities, such as resource management and grid monitoring, the most important feature in the context of the present research. Ganglia monitoring utilises daemon processes, called *gmetad* or *gmond*, which serve to collect data about grid nodes. Ganglia monitoring capabilities can be extended with the *Ganglia metric tool* (*gmetric*), which enables monitoring of arbitrary host metrics by expanding the core set of metrics measured by *gmond* by default. Simply put, users can develop their own metrics and deploy them within the monitored grid system to extend the basic monitoring capabilities. Another important component is the *Job Submission Handler*, which is responsible for resource management and job scheduling. This component is capable of checking whether resources are available, and if users are authorised to access them. Together with the other standard metrics and extended custom metrics, it is possible to extract statistical data for each user, such as the number of allocated nodes, CPU/memory usage, network traffic, user IP address, and so on, from the job submission handler, and deliver this to Ganglia in an adapted format.

- *The integration layer* is where GridSPMS is located. The proposed system is integrated with Ganglia and is able to receive relevant data from it. GridSPMS implements mechanisms for detecting a security incident, and consequently for responding to it. It consists of four main components: Data Collection (DC), Breach Detection Engine (BDE), Security Policy Database (SPD), and Incident Response Module (IRM). These will be described in more detail in the next section.
- *The client application layer* constitutes interfaces through which GridSPMS users can interact with the system. These include an Application Programmable Interface (API), Command Line Interface (CLI), and Web-based Graphical User Interface (GUI). Depending on requirements and personal preferences, users can choose any of the interfaces to control and interact with GridSPMS. Moreover, the interfaces also serve to extend the system functionality and integrate custom software systems (e.g., science, finance, and engineering applications) – that is, to enable the core system with perform monitoring security-related activities in the context of various domain-specific applications.

3.2 GridSPMS components

Having presented a high-level overview of GridSPMS, this section will describe in more detail the internal organisation and individual components of the system.

As previously explained, GridSPMS is built onto Ganglia, which provides the basic tools for resource monitoring in computational grids. From this perspective, it can be seen as an extension of Ganglia, one that complements the existing functionality by providing support for the security dimension of grid computing. The diagram in Figure 2 schematically depicts the main components of GridSPMS, and their interaction with each other. There are 12 main components that can be distinguished, which will now be discussed in more detail in turn.

It is also worth mentioning that the presented component architecture was inspired by the Common Component Architecture (CCA) (Armstrong *et al*., 1999), a reference model that suggests complex computing systems must be designed and implemented in a modular manner. This architecture has been widely used in the design of cluster and grid systems, where individual components are the basic building blocks used to construct a complex system. According to this reference architecture, complex computer systems have to be broken down into loosely-coupled and self-contained elements, which are responsible for implementing certain parts of the overall functionality. Such an approach allows for targeted modifications without affecting the whole system, and therefore enables seamless and timely maintenance.

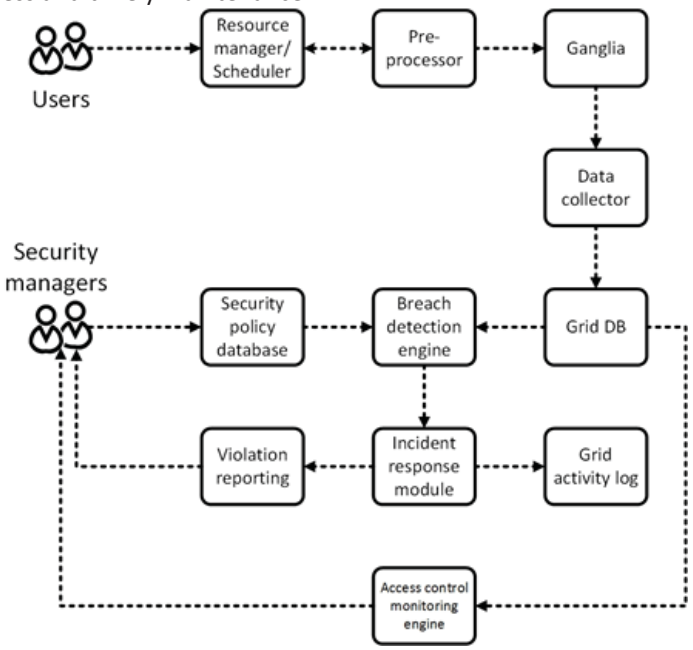


Figure 1: Components of the various layers constituting GridSMPS:

- *The resource manager/scheduler* is part of the core functionality of contemporary grid platforms, and is responsible for i) receiving task requirements from the user; ii) allocating and provisioning appropriate resources to the user; iii) scheduling application jobs and optimising resource utilisation; and iv) monitoring

job execution processes. All of these activities require that the user is authorised to access and use the corresponding grid resources.

- *The pre-processor* is responsible for collecting statistical user data, including jobs, allocated nodes, CPU/memory utilisation, and user IP addresses, which helps to determine users' geographical locations. The pre-processor is also responsible for sending the input information to Ganglia via the *gmetric* tool.
- *Ganglia* typically includes one or more instances of *gmetad* daemon processes, which are responsible for collecting data pertaining to managed grid clusters and nodes. The server running a *gmetad* daemon process continually sends XML packets over the TCP connection; these packets contain information about the current state of individual nodes and clusters within grids. Depending on the types of *gmond* or *gmetad* daemon processes, which are responsible for the actual extraction of data, the XML data will include various fields related to the actual performance of a particular node or cluster, and user behaviour. Ganglia is also designed to be extendible, meaning users are also free to develop and integrate their own extensions to the core system – a feature which has proven helpful in the context of the present research. In particular, Ganglia's support for customised extensions made it possible to enhance it with security-related data probes and metrics, and integrate it with GridSPMS. For example, for tracking the current geographical location of a user so as to decide whether an access should be granted, a Ganglia extension makes it possible to extract and send user geo-location data as an XML packet for analysis by GridSPMS.
- *The data collector* is a component that periodically polls and receives XML packets sent over the network from Ganglia, and parses this information. Depending on predefined rules, it filters and aggregates the obtained information, and stores it in the grid database. At this stage, security concerns begin to play an important role, as the data collector now looks specifically for and at data related to potential security policy breaches. In this sense, the data collector can be considered a mediator between the core Ganglia component and GridSPMS. For example, the data collector can be configured to focus only on geolocation data – it will then ignore the rest of the data (which is considered as 'noise' in the given context) and send the relevant information to the grid database.
- *The grid database* is a standard relational MySQL database, which is responsible for storing data received by the data collector from Ganglia. The remaining data is then available to other components of GridSPMS, such as the BDE and ACME, and administrative tasks can be carried out via the standard SQL queries. Given potentially large amounts of data to be collected and processed in the context of a complex grid ecosystem consisting of hundreds of nodes, it is essential that the supporting database is capable of coping with such a workload. Moreover, in the presence of rapidly changing and varying data, the traditional relational approach to data storage might be insufficient, calling for the emerging Big Data solutions, such as NoSQL databases and stream processing approaches (Sagiroglu and Sinanc, 2013).
- *The breach detection engine* (BDE) is a core analysis component of GridSPMS that queries data from the database and matches it against a set of pre-defined security policies, with the aim of detecting potential security breaches. Situations in which the collected data fails to satisfy the policy constraints are classified as security breaches, and any security incident must be reported, and in some cases responses should be taken. An important issue to be considered in this respect is the type of language to be employed to define the security policies and an associated policy enforcement component. Admittedly, to facilitate a modular and loosely coupled approach, it is essential to use a declarative programming language – that is, a language which enables policy definition to be separated from the actual programming code and the policy enforcement mechanism. With such an approach, it is possible to add new policies and modify existing ones in an agile and seamless manner, without interfering with the main programming code.
- *The security policy database* stores rules, regulating security-related activities within the grid system, and is used by the BDE. It is assumed that security policies are designed and implemented by policy managers, human administrators and domain specialists, who also act as supervisors when responses to incidents are reported (passive responses) or executed (active responses). Unlike the grid database, the security policy database is not expected to handle large amounts of rapidly changing data, and therefore can remain a simple relational database or even a text file.
- *The incident response module* serves to report, manage and document all detected grid security incidents. In case of emergency, for instance if a security breach is detected, the incident response module will first notify the human administrator, and may then take certain automatic actions (thus representing an active response mechanism).

- *The violation reporting component* is a simple passive response module, the only responsibility of which is to notify the human administrator of the detected security breach via one of the available communication channels, for example in an e-mail, a sound signal, or a pop-up window in the administration panel. A notification message contains information about the violation, including detection time, potential cause of violation, and the user associated with the incident.
- *The access control monitoring engine* (ACME) is the component responsible for collecting data relating to system access. It constantly monitors user activity, such as login attempts, user locations and IP addresses, and manages user accounts and access rights, and so on. From the security management point of view, this component is seen as the key source of information to be processed by the BDE, as it keeps track of all the potentially critical activities taking place within the grid environment.
- *The grid activity log* is where all the activities taking place within the managed grid environment are recorded and stored. The storage period may range from few days to several years, depending on the specific requirements. This is an auxiliary component, which is required for the 'post-mortem' analysis of security incidents and system debugging.

4. Conclusion

As discussed earlier in this paper, grid computing systems are complex and dynamic environments requiring appropriate automated management mechanisms, in order to enable stable and reliable operation of the whole grid ecosystem. The research community has responded to this need by proposing a number of monitoring frameworks, which serve to collect data at various levels to support decision-making and management activities within grids. However, the existing solutions appear to include limited support for the collection of security-related data and the enforcement of appropriate security policies and constraints in this respect. With the emergence of mobile grid computing, an increasingly important role of network connections, and a growing number of users remotely accessing computational resources from various locations, grid systems are no longer seen as localised and isolated ecosystems, but are becoming more open and distributed. In this light, it is increasingly important to enable monitoring frameworks with the capacity to collect security-related data and check whether certain security constraints are being complied with; in other words, to monitor and enforce security policies in grids.

As such, this paper presents the authors' current work in progress, which aims at creating an efficient solution to monitor the security of grid systems. The proposed GridSPMS is an extension of the Ganglia framework, and attempts to build on its existing efficient capabilities for monitoring various types of data within grids, to capture additional security-related metrics and match these against a set of security policies.

This paper has described the proposed architecture in a top-down fashion, beginning with a high-level overview of the four main levels contained within GridSPMS. It then focused on the actual internal organisation of the framework, and described its main functional components. More specifically, the way in which the core functionality of Ganglia could be extended to incorporate the collection of security-related data, such as a user's geo-location, to support compliance checking against a set of security policies was explained. The resulting system is expected to be capable of detecting and reporting on various security-related accidents, an increasingly important feature in the context of emerging mobile grids.

To sum up, the present paper describes a work in progress, which has thus far primarily focused on developing the architectural design of GridSPMS. The next step is to consider the implementation of a prototype version of the proposed system, integrate it with Ganglia and validate it in an existing grid environment.

References

Armstrong, R., Gannon, D., Geist, A., Keahey, K., Kohn, S., McInnes, L., Parker, S., and Smolinski, B. (1999) 'Toward a common component architecture for high-performance scientific computing', *Proceedings of The Eighth International Symposium on High Performance Distributed Computing,* pp. 115–124.

Bessis, N., Asimakopoulou, E., French, T., Norrington, P. and Xhafa, F. (2010) 'The big picture, from grids and clouds to crowds: a data collective computational intelligence case proposal for managing disasters', *Proceedings of 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC),* IEEE, pp. 351–356.

Bichhawat, A. and Joshi, R. C. (2010) 'A Survey on Issues in Mobile Grid Computing', *International Journal of Recent Trends in Engineering and Technology*, Vol. 4, No. 2, pp. 15–19.

Cocotas, A. (2012) 'Smartphone Market Forecast: Sales Will Exceed 1.5 Billion Units A Year By 2016', *Business Insider Australia*, 1 March, [Online] available from: http://www.businessinsider.com.au/smartphone-market-forecast-sales-will-exceed-15-billion-units-a-year-by-2016-2012-2 [accessed: 21 November 2015].

Dunnewijk, T. and Hultén, S. (2007) 'A brief history of mobile communication in Europe', *Telematics and Informatics*, Vol. 24, No. 3, pp. 164–179.

Finneran, M. (2013) 'Research: 2013 State Of Mobile Security', *Information Week Reports*, [Online] available from: http://reports.informationweek.com/abstract/18/10935/Mobility-Wireless/Research:-2013-State-Of-Mobile-Security.html [accessed: 21 November 2015].

Foster, I. and Kesselman, C. (Eds.) (1999) *The Grid: Blueprint for a New Computing Infrastructure*, San Francisco, CA: Morgan Kaufmann Publishers Inc.

Foster, I., Kesselman, C., Tsudik, G. and Tuecke, S. (1998) 'A security architecture for computational grids', *Proceedings of the 5th ACM Conference on Computer and Communications Security*, ACM, pp. 83–92.

Foster, I., Zhao, Y., Raicu, I. and Lu, S. (2008) 'Cloud computing and grid computing 360-degree compared', *Grid Computing Environments Workshop, 2008. GCE'08*, IEEE, pp. 1–10.

Ghosal, A. (2015) 'Tim Cook says the PC is dead, but Apple still makes computers', *The Next Web*, [Online] available from: http://thenextweb.com/insider/2015/11/10/tim-cook-says-the-pc-is-dead-but-apples-still-making-desktops-and-laptops/ [accessed: 21 November 2015].

Katsaros, K. and Polyzos, G. C. (2008) 'Mobility-Aware Grid Computing', *The Encyclopedia of Information Science and Technology*, 2nd ed, Hershey, PA: IGI Group.

Litke, A., Skoutas, D. and Varvarigou, T. (2004) 'Mobile grid computing: Changes and challenges of resource management in a mobile grid environment', *Proceedings of 5th International Conference on Practical Aspects of Knowledge Management,* PAKM.

Sagiroglu, S. and Sinanc, D. (2013) 'Big data: A review', Proceedings of 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013. IEEE, pp. 42–47.

Suwan, A., Siewe, F. and Abwnawar, N. (2016) 'Towards Monitoring Security Policies in Grid Computing: a Survey', to appear in *Proceedings of IEEE Technically Sponsored SAI Computing Conference 2016*.

Applied web Traffic Analysis for Numerical Encoding of SQL Injection Attack Features

Solomon Ogbomon Uwagbole, William Buchanan and Lu Fan
Edinburgh Napier University, Edinburgh, UK
s.uwagbole@napier.ac.uk
b.buchanan@napier.ac.uk
l.fan@napier.ac.uk

Abstract: SQL Injection Attack (SQLIA) remains a technique used by a computer network intruder to pilfer an organisation's confidential data. This is done by an intruder re-crafting web form's input and query strings used in web requests with malicious intent to compromise the security of an organisation's confidential data stored at the back-end database. The database is the most valuable data source, and thus, intruders are unrelenting in constantly evolving new techniques to bypass the signature's solutions currently provided in Web Application Firewalls (WAF) to mitigate SQLIA. There is therefore a need for an automated scalable methodology in the pre-processing of SQLIA features fit for a supervised learning model. However, obtaining a ready-made scalable dataset that is feature engineered with numerical attributes dataset items to train Artificial Neural Network (ANN) and Machine Leaning (ML) models is a known issue in applying artificial intelligence to effectively address ever evolving novel SQLIA signatures. This proposed approach applies numerical attributes encoding ontology to encode features (both legitimate web requests and SQLIA) to numerical data items as to extract scalable dataset for input to a supervised learning model in moving towards a ML SQLIA detection and prevention model. In numerical attributes encoding of features, the proposed model explores a hybrid of static and dynamic pattern matching by implementing a Non-Deterministic Finite Automaton (NFA). This combined with proxy and SQL parser Application Programming Interface (API) to intercept and parse web requests in transition to the back-end database. In developing a solution to address SQLIA, this model allows processed web requests at the proxy deemed to contain injected query string to be excluded from reaching the target back-end database. This paper is intended for evaluating the performance metrics of a dataset obtained by numerical encoding of features ontology in Microsoft Azure Machine Learning (MAML) studio using Two-Class Support Vector Machines (TCSVM) binary classifier. This methodology then forms the subject of the empirical evaluation.

Keywords: SQL injection, SQLIA, numerical encoding, input neurons, Azure Machine Learning, training data

1. Introduction

Continuous innovations in internet applications have seen an astronomical growth of data with trending research areas of big data and the Internet of Things (IoT). Mining of large data to address security vulnerabilities are beyond the limitation of signature approach but an alternative machine learning approach provides a solution. Over the years, there are ongoing issues in securing web driven applications. This is evidently demonstrated by the continuous and intrusive attacks originating from many hacking groups including (but not limited to) governments, lone wolf (Khandelwal, 2015), Anonymous and LulzSec rogue groups (Schwartz, 2011; Schone et al., 2014). A hacker, or an intruder, is an individual or group that breaches the security of a computer system by employing an array of techniques to disrupt and pilfer confidential data. A typical method used to steal confidential data is by SQLIA which often leaves any signature driven Web Application Firewalls (WAF) (Appelt et al., 2015) playing catch-up every time there are new attack signatures. Machine learning approaches are able to effectively protect against novel signatures by classifying new attack signatures not trained for as unknown, thereby dropping or referring such requests at the interim.

Current SQLIA research areas are lacking in ready-made robust dataset samples with numeric encoded features. The few non-standard sample files that exist would normally contain unprocessed strings of repeating features of the variations that exist within SQLIA types. The scheme presented here provides a technique with just a regular expression to numerically encode features from any file containing SQLIA features. It also includes a full implementation on how you would extract the dataset from real-time web traffic and deployment to MAML studio to train a supervised learning model implementing TCSVM (Microsoft Azure, 2016). The proposed model is built on MAML (Microsoft Azure, n.d.). The methodology includes: extraction of dataset attributes items and labelling; classification of SQLIA features and validation of the supervised learning model (Kotsiantis, 2007). The model is then exposed as a web service in ongoing SQLIA detection and prevention. Though this proposed model simultaneously delivers a self-contained SQLIA detection and prevention solution, it is intended for evaluating the fitness of dataset extracted by numerical encoding of features ontology using TCSVM statistical model binary classifier. This is the subject of the empirical evaluation in Section 5.

The paper is laid out in six sections ending with a conclusion and future work. Section 2 covers background and theory (attack intent, injection mechanism and SQL Types); Section 3 is focused on related work; with Sections 4 and 5 detailing the features encoding and results.

2. Background theory

The proposed model implements a traditional NFA (Rabin & Scott, 1959; MSDN, n.d.) to match patterns of injection mechanisms and SQLIA types which are then encoded to numerical dataset items. Dataset are input for ANN and ML and this contains numerical attributes of independent x-variables(predictor) and y-dependent variable (what to predict or labelled attribute). Legitimate web requests are expected valid requests that a monitored web application will generate. There are RegEx patterns that validate SQL parsed assembled web requests in-transition at the proxy. The use of proxy and SQL parser in this model means that during backhaul, web traffic can be laid bare and analysed for the attack intent, injection mechanism and SQLIA type. Vulnerable web applications can be susceptible to the following injection mechanisms: injection through web forms; cookies; operating system server instrumentations; and Trojan horse used in second-order attack. The seven notable SQLIA types include: Tautology; Invalid/Logical Incorrect; Union; Piggy-backed; Store procedure; Time-based; and Alternate encoding obfuscation (Halfond et al., 2008).

These notable SQLIA types have derivations within a SQLIA type which can best be described as an intruder tweaking a known attack signature to evade detection e.g. a SQL injection tautological attack of 1=1 can also be written as 1<1, ‘a’=’a’ etc. to achieve the same attack. It is these derivations that exist within a SQLIA type that account for numerous SQLIA features in any large sample dataset populated with repeated strings. This paper proposes a scheme to account for these derivations with the random risk attribute to account for the fact that there are many derivations within a SQLIA type. These random decimal values provide a way to derive large dataset items of both legitimate web requests and SQLIA features (Uwagbole et al., 2016).An intruder would normally first establish a technique to use in SQLIA by probing the target website with a series of trial runs to determine the hotspot and method that best suits the intended target. It may also deploy a combination of SQLIA types for a successful attack. As an example a normal web request will have a query string of http://bsid/bsid/Data Page.aspx?LoginName=bob&Password=@bob which is evaluated in SQL parser to SELECT loginName, password FROM tblUser WHERE loginName='bob' AND password ='@bob'. This legitimate web request can be injected to any of the SQLIA types discussed below. The scheme presented here for encoding numerical attributes from features of strings as detailed in Section 4 is fully replicable with basic background experience in .NET C# and R scripting using open source software of fiddler proxy (Lawrence, n.d.), SQL Script Dom Parser API (MSDN, n.d.), RegEx (MSDN, n.d.) and MAML studio.

Throughout this section simple queries examples (not exhaustive) are presented in the tables to illustrate the different SQLIA types with red flag alerts being matched in the encoding of features into numerical data. The section below provides a high-level overview of the SQLIA types which forms the bases for the pattern matching of both legitimate web requests and SQLIA payload that are encoded into numerical data. The layout of the tables below has: attack intent; web request query string; parsed transact Structured Query Language (tSQL) and an indication (pattern) of red flags being pattern matched.

2.1 Tautology

This SQL injection type of attack is carried out by injecting vulnerable sites with query strings that are altered

in transition to retrieve data from the database. The classical example is by assigning altered strings to the WHERE clause after incorrectly terminating the query with a single quote (‘) or with a tautology statement like *OR 1=1* as shown in Table 1. The outcome is always parsed by SQL parser to be true meaning that the security validation of login credentials will be bypassed to retrieve all the records at the affected back-end database table.

Table 1: Tautology queries

Attack intent	Injectable hotspots, circumventing authentication, extracting data	Pattern
Query string	http://localhost/bsid/DataPage.aspx?LoginName=bob'OR%201=1-- &Password=	`--,1=1,1<1 ‘a’=’a’ etc.
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE loginName= 'bob' OR 1=1--	

2.2 Invalid/logical incorrect query

This is often used to probe the vulnerabilities of the target prior to an attack. The information gained during an initial attack is useful to the intruder to determine what form of further attack to carry out on the target. The example query string in Table 2 will return a divide by zero error if the login account is not *sa* or using *dbo* schema.

Table 2: Logical incorrect queries

Attack intent	Detecting vulnerabilities, extracting data, database finger-printing	Pattern
Query string	http://localhost/bsid/DataPage.aspx?LoginName=';IF((SELECT%20user)%20=%20 'sa'%20OR%20(SELECT%20user)%20=%20'dbo')%20SELECT%201%20ELSE%20SEL ECT%201/0;--2%80%99&Password=	Duplicate SELECT IF, ELSE sa, dbo,1/0
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE loginName=’; IF ((SELECT user) = 'sa' OR (SELECT user) = 'dbo') SELECT 1 ELSE SELECT 1/0; --	

2.3 Union

This attack exploits the UNION command ability to query multiple database tables to retrieve confidential data far beyond the tables used in the web application as shown in Table 3. There are intermediate steps of using ORDER BY to get column names (sqlinjection, 2016).

Table 3: Union queries

Attack intent	Data extraction and bypassing authentication	Pattern
Query string	http://localhost/bsid/DataPage.aspx?LoginName=&Password=UNION %20SELECT%20CreditNo,%20CustAddress%20from%20tblCreditinfo	‘ UNION SELECT
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE loginName=' ' UNION SELECT CreditNo, CustAddress from tblCreditinfo	

2.4 Piggy-backed

The intruder exploits the semicolon (;) to append a valid SQL statement to further SQLIA. In Table 4, an intruder uses another SQLIA type (sqlinjection, 2016) to obtain table name and then the SQL statement is piggy-backed to run SELECT * from tblCreditinfo in order to gain unauthorised credit card information.

Table 4: Piggy-backed queries

Attack intent	Extracting data beyond the scope of the web application and executing commands	Pattern
Query string	http://localhost/bsid/DataPage.aspx?LoginName=;%20SELECT%20*%20from %20tblCreditInfo%20--	‘ ;
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE LoginName=''; SELECT * from CreditCardInfo --	

2.5 Store procedure

The intruder elicits database information by exploiting xp_cmdshell if enabled to trigger a Trojan horse file for malicious attack. Also stored procedures are vulnerable to privilege escalation, buffer overflows, and even manipulated to gain elevated permission access to perform operating system wide operations (Halfond et al., 2008). In the example query presented in Table 5, an attacker runs xp-cmdshell against a spurious text files loaded in the target to circumvent the security in the database.

Table 5: Store procedure queries

Attack intent	Privilege escalation, remote command, performing denial of service	Pattern
Query string	http//bsid/bsid/login.aspx? LoginName=’ ’;exec xp_cmdshell 'attrib "c:\test\spuriousfile.vbs" +r'	;, Exec, xp_cmdshell Attrib c:\test\ spuriousfile.vbs " +r
Parsed tSQL	SELECT * FROM tblUser where loginname =''; exec xp_cmdshell 'attrib "c:\test\ spuriousfile.vbs " +r'	

2.6 Time-based

This is a timed delayed type of SQLIA where an intruder probes a site to elicit a response after a period of time. The query shown in Table 6 will display a response after ten seconds that provides an intruder with information to further more attacks.

Table 6: Time based queries

Attack intent	Probing of injectable hotspots and database schema	Pattern
Query string	http://localhost/bsid/DataPage.aspx?LoginName=';%20waitfor%20delay%20'00:00:10'--&Password=	' waitfor delay '00:00:10' --
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE loginName=' '; waitfor delay '00:00:10'--	

2.7 Alternate encoding obfuscation

A combination of escaped-encoding and Unicode character which the computer systems interpret as normal without distinction from intended obfuscation; an intruder can circumvent solutions being provided in pattern matching of SQLIA as it becomes unreadable as shown in Table 7.

Table 7: Alternate encoding obfuscation

Attack intent	Obfuscating data to evade detection	Pattern
Query string	http%3A%2F%2Flocalhost%2Fbsid%2FDataPage.aspx%3F%0ALoginName%3Dbob%27OR%25201%3D1--26Password%3D%0A	% Hex values Numeric values --
Parsed tSQL	SELECT loginName, password FROM tblUser WHERE loginName=' '; waitfor delay '00:00:10'--	

Not limited to the above injection mechanisms and SQLIA types which are the subject of numerical encoding of features in this paper; there are other forms of pattern evading techniques like the use of comments, whitespace, character casing and encryption that are exploited by an intruder in SQLIA. Section 3 discusses related work.

3. Related work

Whilst it is acknowledged that there are existing foundational works (Boyd & Keromytis, 2004; Gould et al., 2004; Buehrer et al., 2005; Halfond & Orso, 2005) that share similarities with the approach used, these similarities do not extend to the proposed scheme to encode numerical attributes as presented in this paper. In recent work (Uwagbole et al., 2016) introducing the numerical encoding of features ontology to obtain dataset attributes was applied to ANN and statistical ML models implemented using Two-Class Averaged Perceptron (TCAP) and Two-Class Logistic Regression (TCLR) classifier respectively that gave a prediction rate of Area Under Curve (AUC) of 0.914 (91.4%). In this paper, it presents a full implementation on how you would extract the dataset from real-time web traffic directly from MAML and trained a supervised learning model implementing TCSVM classifier with an improved performance results of AUC of 0.944 (94.4%) shown in confusion matrix (performance metrics) in Section 5.

Most recent works are derivations of foundational works on SQL injection that often benchmark detection rates against these earlier research foundational works (Buehrer et al., 2005; Wu et al., 2015), (Halfond & Orso, 2005; Wang et al., 2015) devoid of today's big data challenges with growing internet data. As intruders become ever smarter in developing novel techniques to tweak known SQLIA types with whitespaces, character casing, comments, encryption and encoding, so it becomes evident that the static signature methods (Boyd & Keromytis, 2004; Kar et al., 2015) lack the ability to cope, having to continuously create new signatures. Exploring an alternative machine learning approach is a direction towards a better prediction of scalable processing demand of growing internet data traffic.

SQLProb (Liu et al., 2009) uses proxy in their approach, as does this proposed model, but this proposed model goes further in applying supervised learning in prediction of true positives and negatives as against genetic algorithm. SQL parsing tree (Buehrer et al., 2005) and CANDID (Bisht et al., 2010) are centred on code analysis algorithms for detection of SQLIA. The benefit asserted by the authors of the approach is that it can be retrofitted, but as these methods lack pattern matching the approach will not be functional in today's high volumes of web traffic. Though AMNESIA (Halfond & Orso, 2005) is a hybrid of static and dynamic approach that uses NFA as does this paper; in the AMNESIA experiment, a Java based NFA implementation of SQLIA

pattern searches were used solely in SQLIA detection and prevention but in the approach presented here, the matched pattern is numerically encoded into dataset items that are fed into ANN and ML implemented on MAML platform. SQLrand (Boyd & Keromytis, 2004) is another related work that employs proxy and parser. Although the authors claim the methodology to have a negligible impact on performance, it is unlikely that it would be scalable in today's large data driven web traffic due to their approach being geared towards signatures as against pattern matching employed in AMNESIA. JDBC Checker (Gould et al., 2004) is a static approach, which although it explores finite state automaton that is also used in this model, it lacks proxy to backhaul web traffic as to lay bare web requests for thorough analysis.

4. Features encoding

Section 4.1 discusses features encoding to numerical attributes items destined as input dataset to a supervised learning model implementing TCSVM while Section 4.2 is a high level overview of the full implementation steps on MAML studio.

4.1 Numerical attributes extraction from features steps

The dataset input to the supervised learning model are features numerically encoded from legitimate web requests, injection mechanisms and known SQLIA types. The attributes (predictors) are scaled down in this paper for simplicity, but the approach can be replicated for as many attributes that are desired and attributed to SQLIA behaviour patterns. Table 8 shows the semantics of how the numeric attributes data items are abstracted.

Table 9 contains numerical attributes data items that are extracted as detailed in Table 8 which has the dataset items of both numerical encoded features of normal and SQLIA:

- Sitypes *p* of recognised patterns p $\{w_{0\ldots} w_n\}$
- Sidetermination *v* cross validated feature types v $\{w_{0\ldots} w_n\}$ using a method of NFA backtracking (MSDN, n.d.)
- Rndrisk *r* is randomised values to account for the variations within a feature $r_0 \ldots r_n$.

Siriskfactor *l* is the likeliness of a feature being a risk factor which can either be -1 likeliness or 1 for remote likeliness. This is collated from the predictor independent variables (x-attributes) of *p, d, r*. The dependent y-variable or what to predict (commonly known as a labelled class) is inferred from *l* which could be a possible 1 for SQLIA or 0 for normal as shown in Table 8 with steps detailing how these values were extrapolated. The scheme can be replicated with any pattern or text processing tool like RegEx to assign number range (numeric attributes data items). However, a sum of the attributes needs to be less than or greater than the value set for the threshold in likeliness of it being normal or suspect.

Table 8: Numerical encoding of features algorithm *(Uwagbole et al., 2016)*

```
Extracting the primary independent predictors
        Get matched patterns (p) of SQLIA and legitimate payload using NFA(RegEx)
                if feature matched a static pattern
                assign a numeric value from 1 to 9
        Validate (v) matched pattern(p)
                randomly assign values 0.01 to 0.09
        Randomization (r) to account for the variations within injection mechanisms and SQL types
                randomly assigned a decimal value less than 0.01
                computed against web requests total count
Calculating the likeliness of being normal (n) threshold
                Sum p + v + r and take the minimal from the set of values above 9
                 If   n >= 9
                      Then
                                Normal = 1
                  Else
                                Suspect = -1
    3.  Calculating y –variables or what to predict
                 If normal(n)
                       Then
                            0
                 Else
                            1
```

Table 9: Encoded numerical attributes data items for input ML

Sitype(p)	sidetermination (v)	Rndrisk (r)	Siriskfactor (l)	Siclass
$p\{w_0\}$	$v\{w_0\}$	r_0	l_0	y_0
1	*0.04*	*0.389420*	*-1*	*1*
9	*0.09*	*0.142830*	*1*	*0*
$p\{w_n\}$	$v\{w_n\}$	r_n	l_n	y_n

4.2 The implementation steps on MAML studio

Figure 1 is the MAML experiment screen capture that is carried out for the proposed model. The steps include: streaming the extracted numerical dataset items into MAML studio; classification of SQLIA features; validation of supervised learning model which is then exposed as web services in ongoing detection and prevention. Below are the nine key steps:

- The web traffic is backhauled at proxy to analyse web requests for injection mechanisms, legitimate web requests and SQLIA features.
- The web requests payload in-transition to the back-end database are assembled to full queries and parsed by a SQL parser API at the proxy.
- The payload is subjected to pattern matching employing a traditional NFA approach (RegEx) while at the same time numerically encoding both matched and unknown patterns in web requests.
- The numerical encoded dataset is streamed into the MAML studio by R scripting call to traditional NFA implemented using Microsoft RegEx patterns matching at the proxy.
- This is followed by scrubbing of the missing data and column casting as the input dataset is feature engineered with missing values to achieve a better confusion matrix that can accurately predict SQLIA. This helps remove over-fitting from the binary classification.
- Next step is logistic normalisation which is the rescaling of the numeric data as to constrain the values (Microsoft Azure, 2015b) for a good fitting of the numeric data into the statistical trained model.
- Next is the splitting of data between training and testing data. A repeated training of the statistical model with different splits ratios found using 80% training data to 20 % test data that gave the highest prediction rates shown in Section 5 confusion matrix (figure 2).
- The supervised learning model was trained using TCSVM classifier algorithm. It was scored and evaluated to establish how well the prediction of the supervised learning classification model is performing. This gave AUC value of 0.944 (94.4%) shown in confusion matrix in Section 5 (figure 2).
- Finally, a predictive web service was generated from the trained supervised learning model built using MAML studio to obtain an API code to integrate into web form for ongoing SQLIA prediction.

Figure 1: The MAML studio predictive experiment layout.

Whilst all new web requests from new IP addresses go through the pattern matching process but the established authenticated IP addresses are validated through the web service of the supervised classification trained model for ongoing scalable SQLIA detection and prevention.

5. Evaluation and result

There were four attributes or x-variables (sitype, sidetermination, rndrisk, siriskfactor) with one labelled attribute or y-variable (siclass). ML and ANN approaches need a large dataset numerically encoded from behaviour or patterns of normal and SQLIA features to accurately predict. The numerical attributes encoding ontology also detailed in this paper provides a way to generate as many desired dataset attributes and rows items. There were 59702 rows of attributes data items of which 80% (47762 rows items) were used as training data and a further 20% (11940 rows items) were used as the test data. This test data is the subject of computational statistics in any performance metrics (confusion matrix). There are 1:8 ratios between normal and attack features in the distribution of the attributes data items (rows) and a further outliers of data items with missing values. Through repeated training of the supervised classifier using different normalisation and split ratios, an optimum classification was achieved at normalisation using logistic data transformation method with a split data ratio of 80:20 between training to test data which achieved an AUC value of 0.944 (94.4%) on TCSVM classification model.

The performance metrics or confusion matrix (accuracy, precision, recall and F1 score) from Figure 2 are calculated as follows:

- Accuracy is the proportion of actual true results to the total cases that is calculated as: true positives + true negatives (10446 + 719) / total cases (11940) = 0.935.
- Precision is the proportion of true overall positive results returned by the model that is calculated as: true positives (10446) / true positives + false positives (10446+605) = 0.945.
- Recall is the true positive rate which is fraction of total correct results returned by the model calculated as: true positives (10446) / all positive cases (true positives (10446) + true negatives (170) = 0.984.
- F1 score is a measure of accuracy that balances precision and recall (Microsoft Azure, 2015a) calculated as: 2 *(recall (0.984)*precision(0.945))/(recall(0.984)+ precision(0.945)) = 0.964.

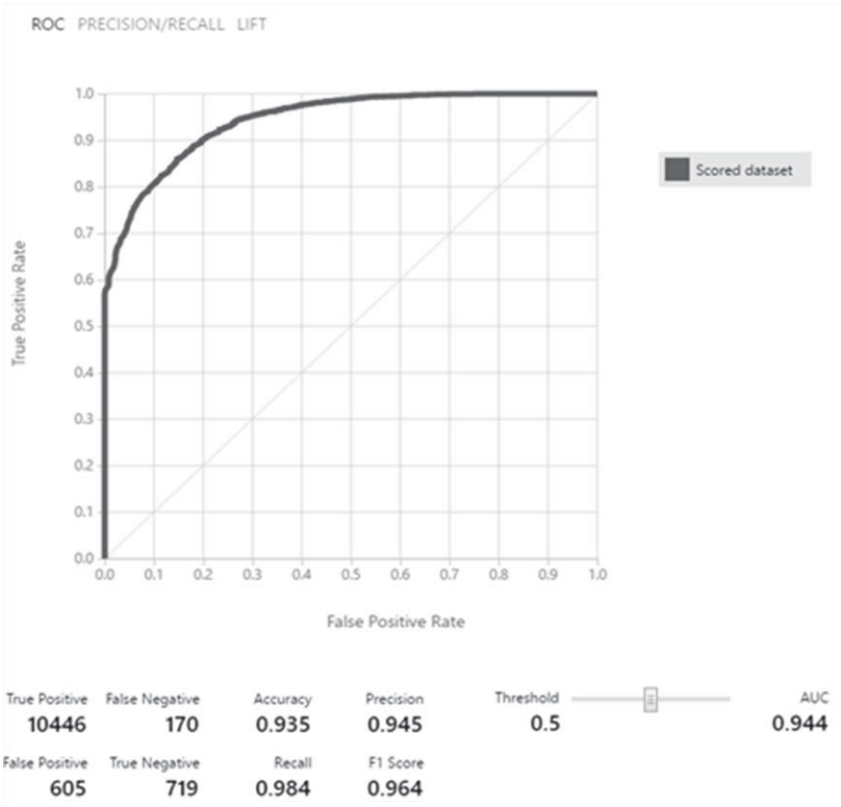


Figure 2: ROC plot and confusion matrix of evaluation results

Receiver Operating Characteristic (ROC) is a graphical plot (de Ruiter, 2015) in Figure 2 using variation in threshold discrimination to illustrate the performance of the binary classifier system implementing TCSVM classifier with a curve towards the upper left corner indicating a better performing model. Area Under the Curve (AUC) or area below the curve in the graph plot is a measure of true positives on the y-axis against false positives on the x-axis. An excellent prediction model is inferred with AUC 0.944 as shown in Figure 2 which indicates the scheme presented will efficiently predict true positives and negatives as required in any effective ML SQLIA detection and prevention model.

6. Conclusion and future work

The work presented in this paper demonstrates the fitness of numerical encoding of web requests primed as dataset (detail in Table 8 & 9) to a statistical binary classifier implemented in MAML using TCSVM classifier. The evaluation results presented above in the ROC graph plot and confusion matrix empirically evaluates the proposed scheme to apply ML in real-time for a scalable prediction of SQLIA. This approach is geared towards leveraging recent advancement in the field of artificial intelligence to build a scalable application that can predict SQLIA in web requests with a high degree of accuracy in true positives and negatives.

Whilst an excellent supervised predicting model has been achieved and tested with a further coding to map back the numerical features to interpret string features but a future work is needed to directly interpret the SQLIA strings from the trained supervised learning model.

References

Appelt, D., Nguyen, C.D. & Briand, L. (2015) Behind an application firewall, are we safe from SQL injection attacks? In *2015 IEEE 8th International Conference on Software Testing, Verification and Validation, ICST 2015 - Proceedings*.

Bisht, P., Madhusudan, P. & Venkatakrishnan, V.N. (2010) CANDID: Dynamic candidate evaluations for automatic prevention of SQL injection attacks. *ACM Trans. Inf. Syst. Secur.*, 13(2), pp.14:1–14:39. Available at: http://doi.acm.org/10.1145/1698750.1698754.

Boyd, S.W. & Keromytis, A.D. (2004) SQLrand: Preventing SQL injection attacks. In *Applied Cryptography and Network Security*. pp. 292–302. Available at: http://link.springer.com/chapter/10.1007/978-3-540-24852-1_21.

Buehrer, G.T., Weide, B.W. & Sivilotti, P.A.G. (2005) Using Parse Tree Validation to Prevent SQL Injection Attacks. In *Proceedings of the 5th international workshop on Software engineering and middleware SEM 05*. p. 106. Available at: http://portal.acm.org/citation.cfm?doid=1108473.1108496.

Gould, C., Su, Z. & Devanbu, P. (2004) JDBC checker: a static analysis tool for SQL/JDBC applications. In *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*. pp. 697–698.

Halfond, W.G.J. & Orso, A. (2005) AMNESIA: Analysis and Monitoring for NEutralizing SQL-injection Attacks. *Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering*, pp.174–183. Available at: http://doi.acm.org/10.1145/1101908.1101935.

Halfond, W.G.J., Viegas, J. & Orso, A. (2008) A Classification of SQL Injection Attacks and Countermeasures. *Preventing Sql Code Injection By Combining Static and Runtime Analysis*, p.53.

Kar, D., Panigrahi, S. & Sundararajan, S. (2015) Sqlidds: SQL injection detection using query transformation and document similarity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 377–390. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84922376264&partnerID=tZOtx3y1.

Khandelwal, S. (2015) Fourth, a 16-year-old Hacker, Arrested over TalkTalk Hack. *The Hacker News*. Available at: http://thehackernews.com/2015/11/talktalk-hacker.html [Accessed January 23, 2016].

Kotsiantis, S.B. (2007) Supervised Machine Learning : A Review of Classification Techniques. *Informatica*, 31, pp.249–268. Available at: http://books.google.com/books?hl=pt-BR&lr=&id=vLiTXDHr_sYC&pgis=1.

Lawrence, E. Fiddler free web debugging proxy. *Telerik*. Available at: http://www.telerik.com/fiddler [Accessed February 11, 2015].

Liu, A. et al. (2009) SQLProb : A Proxy-based Architecture towards Preventing SQL Injection Attacks. *System*, pp.2054–2061. Available at: http://portal.acm.org/citation.cfm?id=1529282.1529737.

Microsoft Azure (2015a) Machine Learning / Evaluate. *MSDN Library*. Available at: http://tinyurl.com/zybaw94 [Accessed February 1, 2016].

Microsoft Azure Microsoft Azure Machine Learning Studio. *Microsoft Azure Machine Learning*. Available at: https://studio.azureml.net/ [Accessed January 25, 2015].

Microsoft Azure (2015b) Normalize Data. *MSDN Library*. Available at: https://msdn.microsoft.com/en-us/library/azure/dn905838.aspx [Accessed January 6, 2016].

Microsoft Azure (2016) Two-Class Support Vector Machine. *MSDN Library*. Available at: https://msdn.microsoft.com/en-us/library/azure/dn905835.aspx [Accessed January 22, 2016].

MSDN Matching Behavior. *MSDN Library*. Available at: https://msdn.microsoft.com/en-us/library/0yzc2yb0(v=vs.100).aspx [Accessed February 3, 2016a].

MSDN Microsoft.SqlServer.TransactSql.ScriptDom Namespace. Available at: https://msdn.microsoft.com/en-us/library/microsoft.sqlserver.transactsql.scriptdom.aspx [Accessed October 25, 2015b].

Rabin, M.O. & Scott, D. (1959) Finite Automata and Their Decision Problems. *IBM Journal of Research and Development*, 3(2), pp.114–125.

de Ruiter, A. (2015) Using ROC plots and the AUC measure in Azure ML | Andreas De Ruiter's BI blog. *MSDN*. Available at: https://blogs.msdn.microsoft.com/andreasderuiter/2015/02/09/using-roc-plots-and-the-auc-measure-in-azure-ml/ [Accessed March 15, 2016].

Schone, M. et al. (2014) War on Anonymous: British Spies Attacked Hackers, Snowden Docs Show. *Nbcnews.Com*. Available at: http://www.nbcnews.com/news/investigations/war-anonymous-british-spies-attacked-hackers-snowden-docs-show-n21361.

Schwartz, M.J. (2011) Sony Hacked Again, 1 Million Passwords Exposed. *InformationWeek*. Available at: http://www.darkreading.com/attacks-and-breaches/sony-hacked-again-1-million-passwords-exposed/d/d-id/1098113?

sqlinjection (2016) SQL Injection Using UNION. *sqlinjection*. Available at: http://www.sqlinjection.net/union/ [Accessed February 3, 2016].

Uwagbole, S., Buchanan, W. & Fan, L. (2016) Numerical Encoding to Tame SQL Injection Attacks. In *IEEE/IFIP DISSECT*. In press.

Wang, Y. et al. (2015) Detecting SQL Vulnerability Attack Based on the Dynamic and Static Analysis Technology. In *2015 IEEE 39th Annual Computer Software and Applications Conference*. IEEE, pp. 604–607. Available at: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=7273432 [Accessed March 20, 2016].

Wu, T.. et al. (2015) Towards SQL injection attacks detection mechanism using parse. In *In Genetic and Evolutionary Computing*. Springer International Publishing, pp. 371–380.

Masters Research paper

Architectural Requirements Specifications for Designing Digital Forensic Applications

Stacey Omeleze and Hein Venter
Information Computer Security Architecture (ICSA) Research Group, Computer Science Department, University of Pretoria, South Africa
someleze@cs.up.ac.za[1]
hventer@cs.up.ac.za[2]

Abstract: The purpose of digital forensic applications is to determine the flow of events as they unfold, before, during or after the alleged incident occurred. The need to identify the cause of an incident, the flow of events preceding the incident as well as prove the consistency of potential evidence recovered from the alleged incident demands a proactive approach to the designing of digital forensic applications employed in the investigation of such incidents. In the field of digital forensics, applications are used to unravel the cause of incidents, especially during acquisition, analysis or examination of potential digital evidence. The frequent updating of digital devices that are susceptible to involvement in digital crime incidents is a huge challenge to digital forensic investigations. To proactively overcome this challenge, this paper proposes an architectural requirement engineering specifications (ARES) process for the design of digital forensic (DF) applications. The proposed ARES process outlines the design process of DF applications using the online neighbourhood watch (ONW) system for its case study. The result shows that in employing the ARES process in designing DF applications, modifiability, pluggability and maintainability features and changes in legal requirements are achieved, thereby accommodating the constant upgrade/changes associated with electronic devices operating systems (OS) or hardware which may sometimes be involved in DF investigations.

Keywords: digital forensics, DF applications, requirements engineering, digital investigation, Df tools, online neighbourhood watch, neighbourhood crime, architectural requirements

1. Introduction

The primary objective of this research is to propose an architectural requirements engineering specifications (ARES) process for designing forensically sound applications. Using the online neighbourhood watch (ONW) system as a case study, the ARES process is illustrated. Whether the DF application is used in conducting an organisational inquiry or criminal investigation, the DF application is used to trace potential digital evidence (PDE) and in most cases it provides the required clues as to what happened. At any given stage in the lifecycle of a DF application, it is used to discover, extract or analyse digital evidence.

The constant change in technology, continuous updates and the emergence of new operating systems (OS) are some of the reasons for incorporating architectural requirements engineering specifications in designing DF applications. These changes and updates are accommodated by providing pluggable features at the design stages of any DF application. Using the ARES process in designing DF applications ensures flexibility, modifiability and pluggability features are embedded in the DF application at its architectural specifications.

In a recent work, the authors proposed the ONW model, which uses the existing functions of mobile devices, that is video camera, voice recording and image capturing functions as a real- time potential digital evidence (PDE) capturing tool to acquire and store PDE of neighbourhood crime in South Africa (Omeleze and Venter, 2014). The stored PDE is made available to law enforcement agents, the judiciary and other stakeholders to be used in criminal related investigations. However, in implementing the ONW system or any other digital forensic application, an adequate requirements specifications process is necessary to effectively achieve the system's objectives.

The ARES process implements a component-by-component based and retraceable design system, where design accuracy and definitive standards are enforced to increase user confidence in a DF application. User confidence as it concerns public understanding of science and technology, while exploiting legal resources to resolve technical controversies in a court of law, are some challenges that using the ARES process deals with. The design processes of a DF application can essentially be presented during a digital forensics presentations and reporting in a court of law (Omeleze and Venter, 2013)(Valjarevic and Venter, 2012). This is to show the place of forensic soundness in the design and development of a DF application. According to Garfinkel et al., (2009) definitive standards are especially critical for DF applications that are employed during legal proceedings and decision

making. Furthermore, many studies have shown that software failure originates from inadequate requirement engineering specifications (Hofmann and Lehner, 2001). These, therefore are reasons to apply requirements specifications specifically for designing DF applications.

The problem this paper addresses therefore, is that there are no easy means of incorporating requirements specifications in designing DF applications. To address this problem, a architectural requirements engineering specifications (ARES) process that is tailored for designing DF applications, is proposed. Subsequently, the proposed ARES process is employed in designing the ONW system as a case study.

The remainder of this paper is structured as follows: Section 2 and 3 provides a background overview of digital forensics and requirements engineering. Section 4 outlines the proposed ARES process for DF applications. Section 5 presents the case study illustration applying the ARES process to design the ONW system, thus identifying the functional and architectural requirements of the system. Section 6 evaluates this paper, with section 7 concluding it.

2. Digital forensics

Digital forensics uses mathematical techniques to solve problems related to incidents involving data recovery, electronic data disputes and storage devices during legal and other related enquiries (Casey, 2011). It is a field that works in parallel with law and science in the identification of patterns of incidents under investigation, using legal and scientific standards to support or refute digital evidence (Watney, 2009)(Cohen, 2009).

Digital forensics employ proactive, reactive or cohesive methods to solve or estimate various incidents. In doing so, various means such as digital forensic tools, applications, digital forensic standards, expert witness testimonies and scientific theories are adopted to achieve the objectives of digital forensic investigations. Digital forensic investigation seeks to obtain justice by employing laws and regulations that govern digital evidence admissibility and demonstrates events related to crimes using digital forensic tools (Garfinkel et al., 2009)(Watney, 2009). However, evidence presented to a court of law may be brought under scrutiny, by questioning the application used in the analysis of evidence, such as whether it adhered to known scientific processes, theories or methods. Such methods could include the Daubert guidelines, examining the known error rate of the application, and proof of consistency, in the form of expert witness (Cohen, 2009)(Watney, 2009).

Furthermore, the success of digital forensic science depends on establishing and following processes that are outlined in the Daubert guidelines (Orofino, 1996). The Daubert guidelines are used by a presiding Judge in a legal proceeding to determine when an expert's scientific testimony is based on reasoning or scientifically valid methodologies. The Daubert guidelines require that the digital forensic process or methodologies must be tested, peer-reviewed, have a level of acceptance in the scientific communities, known potential error rate, with standards and controls. Employing a known process such as software requirement engineering specifications, when designing a DF application, enhances the DF application's validation when such scrutiny arises in any court of law.

3. Requirements engineering

Requirements engineering of a software system focuses on the quality of software products. It takes into account the various aspects of a system's design decisions and the conditions required to address the system's problems (Pohl, 2010). The success of a software system is measured by the extent to which the various conflicting aspects and stakeholders needs are managed, while adhering to the system’s intended purpose. To effectively convey these requirements, proper communication aligned to the applications' functional and architectural requirements and their constraints constitute the foundation for achieving a solid system design (Len Bass, 2012)(Solms, 2012).

In any system design, the requirements engineering processes employed include the identification of the systems' functional requirements (also known as the users' requirements), the architectural requirements and constraints (Pohl, 2010)(Len Bass, 2012). These requirements and the processes employed to realise them are elaborated on in the sections that follow.

3.1 Functional requirements

Functional requirements identify and define the system components, inputs, the behaviour of the inputs and the resulting output at the system's design stages (Len Bass, 2012) (Leffingwell and Widrig, 2003). It focuses on the behavioural requirements that describe the use cases by capturing the role of the system's users and applying their various functions to the system(Pohl, 2010). The identified needs of the users are mapped to the system's architectural requirements.

3.2 Architectural requirements

Architectural requirements are the system's components required to commence its high-level infrastructural design(Solms, 2012) (Len Bass, 2012). These are quality requirements, architectural patterns and architectural strategies that are used to address the core quality requirements of a system and in turn fulfil the users' requirements. Delivering the quality requirements of a system is critical to the success of the system's core objectives.

3.2.1 Quality requirements

The quality requirements are the infrastructural elements that enable the system to meet the stakeholder's concerns. It is a means to concretely align the desired system's objectives to the architectural roadmap using metrics and scales to quantify the system's expected output, as well as, identifying the trade-offs. For example, in fulfilling the flexibility requirements of a system, security may be traded off. Likewise, in realising the performance requirements of a system, reliability may be traded off. Furthermore, some quality requirements (when over looked at the system design stage) may require a complete rebuild of the system to incorporate the forgone a tribute (Solms, 2012). Examples of quality requirements of any system are security, reliability, auditability, pluggability and maintainability.

3.2.2 Architectural patterns

Architectural patterns, also known as architectural styles are reusable solutions that enable innovation to be incorporated at system's design stages. This is by providing infrastructure used to realise various systems' quality requirements (Solms, 2012). Architectural patterns include pipes and filters, microkernel, blackboard, layered and service oriented architecture.

3.2.3 Architectural strategies

Architectural strategies (also known as architectural tactics) are used to concretely address the quality requirements of a system. It specifies how to carry out a design to concretely fulfil a single quality requirement of a system(Len Bass, 2012). This makes trade-off decisions explicit, thereby assisting in deciding on features that best address a quality requirement. Two or more architectural strategies can be used to address one quality requirement.

3.3 Architectural constraints

Architectural constraints are the conditions a ached to a system's requirement specifications (Pohl, 2010). Constraints are mostly imposed by the stakeholders. Constraints can be legal, environmental, economic, technological, or time factors that restrict the development of a system. The system has to identify a way to incorporate its constraints while at the same time meeting the various requirements of all stakeholders.

In summary, employing requirements engineering for the design of DF applications ensures that DF application's architectural requirements, such as their expected output, incremental pluggability and maintainability functions are implemented. Moreover, requirements engineering specifications as they concern digital forensics are necessary especially because of the need to keep pace with the constant advancement in hardware devices and the frequent upgrades in operating systems (OS) used by these devices that are susceptible to digital forensic investigations. In addition to being in tune with technological changes, DF applications must adhere to the current legal standards. The proposed process for designing digital forensic applications takes into account requirements engineering specifications. This process which is termed architectural requirement engineering specifications (ARES) for designing digital forensic applications is presented next.

4. The proposed architectural requirements engineering specifications process for designing digital forensic applications

Having introduced requirements engineering for any software application, this Section proposes an architectural requirements engineering specifications (ARES) process for designing digital forensic applications.

4.1 The architectural requirement engineering specifications process

The ARES process, employs user's requirements to determine the application's architectural specifications to fulfils the system's overall goal.

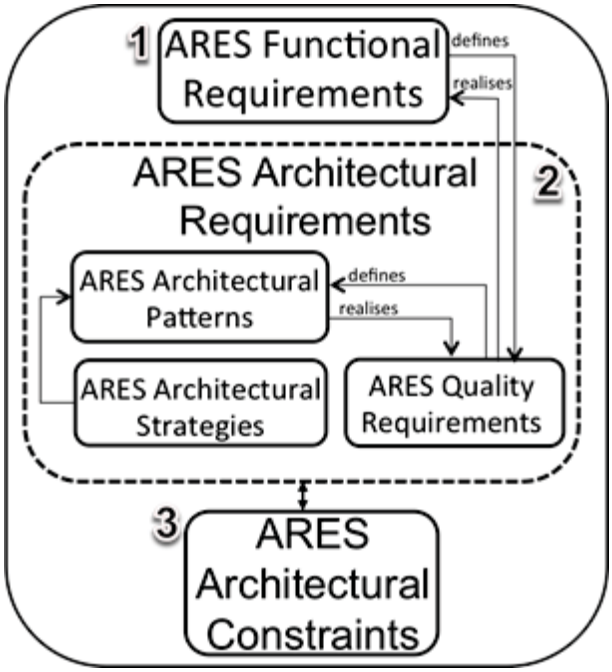


Figure 1: High level view of the proposed ARES process

Figure 1 depicts the high level view of the ARES process which consists of: (i) The ARES DF application functional requirements. (ii) The ARES DF application architectural requirements, which consists of quality requirements, architectural strategies and patterns. (iii) The ARES DF applications constraints. Each of the processes (see Figure 1) are decomposed into lower level functions showing their uses in designing DF application. These processes are discussed in the sections that follow

4.1.1 ARES functional requirements

In using the ARES process to identify the functional requirements of a DF application, the system focuses on identifying the main users' needs of the system as shown in Figure 2.

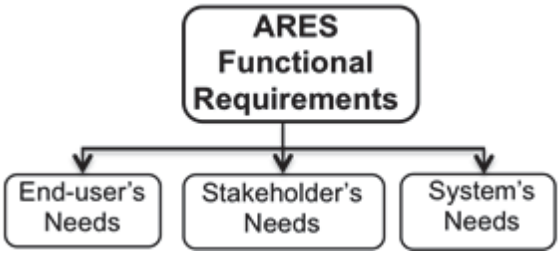


Figure 2: ARES process for functional requirements specifications

Using these identified needs of the system, the functional requirements concerns of the DF application are represented using an use case diagram (Leffingwell and Widrig, 2003). In designing DF applications, there are various stakeholders needs that must be aligned to the needs of all concerned with its design, while ensuring that the requirements of the DF application system are achieved. However, aligning these various needs of the DF application's stakeholders can result in trade-offs that either enhance or impede the performance of the DF

applications. Therefore fulfilling the core requirements is the basic priority of the ARES process. However, there are some needs of the DF applications that are common to all stakeholders, these are termed the core needs. The core needs are where all the stakeholders requirements converge as depicted in Figure 3. The core needs of a DF application cannot be traded-off, rather other requirements of the DF application defer to the core needs (see Figure 3).

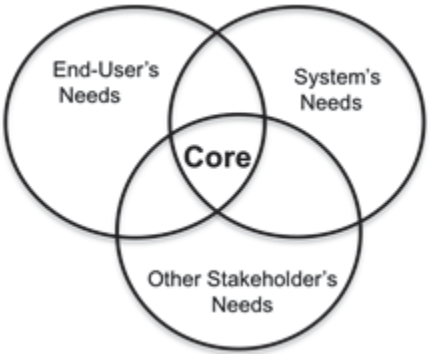


Figure 3: Core users of any DF application

The requirements of all stakeholders across the system are mapped to the functional requirements specifications of a DF application. As depicted in Figure 3 the stakeholders of a DF application consist of the end-users, the system's needs and other stakeholder's needs, with the core as the overlapping convergence point. To identify the functional requirements of any DF application, the authors propose the following process:

(i)Identify the needs of the application and that of the users. These are the unique and elicited activities to be accomplished by the application, as well as the user's expected output as the final product. The identified needs are decomposed to a lower level of granularity, in order to arrive at a concrete process used to achieve the functional requirements, using use cases, sequence or activity diagrams.

(ii) The identified user's requirements must be interpreted to determine the architectural needs that best realise the DF application's functional requirements. This is because architectural requirements are focused on realising the qualities of any system using architectural strategies and architectural patterns as its design baseline.

4.1.2 ARES architectural requirements

The architectural requirements consists of the quality attributes, the architectural patterns and strategies used to realise the DF application's identified functional requirements. As shown in Figure 4, the ARES architectural requirements process consists of identifying the quality requirements, the architectural patterns and the architectural strategies. However, the functional requirements directly determine what the quality requirements of a DF application can be, then the identified quality requirements are realised using architectural patterns and architectural strategies

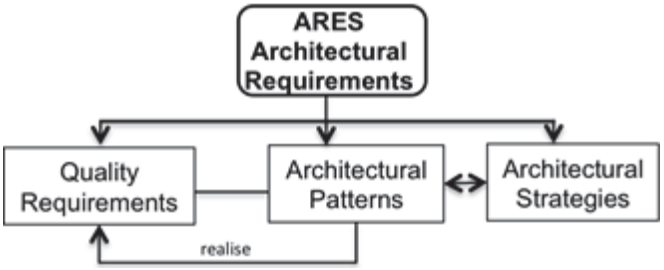


Figure 4: ARES process for architectural requirements design of DF applications

ARES Quality requirements

The approach proposed by the authors towards defining the quality requirements of any DF application includes the following:

(i) Identifying the application's needs using the user's requirements from the functional requirements. (ii) Use the identified users requirements to determine the quality requirements of the application by mapping the

user's needs to the application requirements. (iii) Determining the architectural patterns to be used in order to realise the identified quality requirements. (iv) Determining the architectural strategies that best realise each concrete aspect of the identified quality requirements, while aligning this to the architectural pattern used to fulfil the majority of requirements.

ARES Architectural Patterns

Choosing the best architectural patterns is a critical part of the ARES process. The process proposed by the authors to determine the ARES architectural patterns of a DF application is as follows:

(i) A decision concerning the application's overall architectural responsibility is made. (ii) The architectural pattern that best achieves the application's architectural responsibility - at all levels of granularity is chosen. (iii) The choice of architectural pattern should achieve one or more of the quality requirements of the DF application. For example, if the priority of an application is to ensure security, the microkernel architectural pattern is considered at the first level of granularity, while architectural patterns like the pipes and filters, layered or MVC patterns are employed at the second or third level of granularity.

ARES Architectural Strategies:

Architectural strategies consist of individual requirements that specify the means to address the identified quality requirements of a system. To realise the quality requirements of any DF application using architectural strategies, the authors identified the following process:

(i) Identify the core architectural strategies that best address the architectural patterns that focuses on the overall need of the DF application. For example to address the security requirements of a DF applications, encryption, cryptographic hash and access control are potential strategies to use.

(ii) Select the strategies in-the-order of importance. For example if security is the priority, auditability is put in place before usability requirements are in place. Hence two or more strategies can be used to address one quality requirement.

4.1.3 ARES Architectural Constraints

DF application's architectural constraints are the concerns that are usually identified by the stakeholders that must be considered when the functional and architectural design decisions are reached. As shown in Figure 5 the common constraints of a DF application are jurisdictional, legislative and technological.

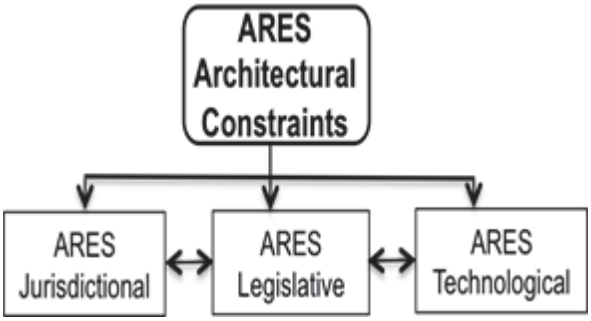


Figure 5: ARES process to define architectural constraints of DF applications

To design a forensically sound application, the authors identified the following process to incorporate architectural constraints.

(i) The jurisdictional constraints span around the acceptable tenancy of data ownership of acquired and stored PDE. The regulations of the data storage domain must comply with the digital data laws of that domain. In South Africa data usage regulations focuses on the ECT, PoPI and RICA Acts (ECT-Act, 2002),(Watney, 2009),(POPI-Act, 2013). The jurisdictional constraints must be taken into account when designing a DF application because each domain has different digital evidence legal requirements, especially as it concerns data retention/ownership.

(ii) In terms of technology the impact of the constant updating of the features of mobile and electronic devices is also a constraint that must be factored in, while designing the application.

(iii) The prescriptions of legislation and the Constitution of the country where the DF application is to be used must be adhered to, especially with regards to laws governing digital evidence admissibility, in a country like South Africa for example (ECT-Act, 2002)(Schwikkard and Van der Merwe, 2009).

4.2 Summary of the architectural requirements engineering specifications process

Since the ARES process has been explained in three distinct phases, this section concludes with an overall summary of the complete process as follows:

(i) Identify the functional requirements, i.e., the users and application's requirements focusing on the application's core requirements.

(ii) Determine the architectural needs of the application using the users and the application's identified functional requirements.

(iii) Identify the core quality requirements based on the user's requirements by mapping the needs of the user to the application's core requirements.

(iv) Ensure the quality requirements of the application are achieved in a technology neutral design using architectural patterns and strategies.

(v) Identify the constraints of the system such as, technological or legislative or jurisdictional, that must be adhered to during the DF application design and development phase.

(vi) Identify the reference architecture that best conforms to the identified quality requirements in-line with the architectural patterns and strategies at various levels of the system's granularity.

5. Applying the ARES process to the ONW system

With the ARES process presented this section demonstrates how the process can be applied to a DF application. As introduced, the ONW model is developed to facilitate the use of mobile devices to capture potential evidence of crime, to assist the law enforcement agencies in evidence gathering and conviction of perpetrators of neighbourhood crime in South Africa or other domains with similar crime situations. In designing the ONW system, the ARES process is employed. The ARES process begins by identifying the functional requirements of the ONW system. Then the functional requirements are used to determine the architectural requirements which comprise of the quality requirements, architectural patterns and strategies. Finally, these architectural patterns and strategies are used to realise the ONW system's main objective. Meanwhile the ONW system's constraints are also incorporated into the application's design decisions. To apply the ARES process in designing the ONW system, it begins with the functional requirements.

5.1 Functional requirements of the ONW system

As depicted in Figure 6, the core functions of the ONW system are: (a) Capture and upload PDE to the ONW repository. (b) Receive Notifications. (c) Validate PDE - by ensuring PDE is forensically sound at capture, upload, and storage. (d) View stored PDE (e) Download and Validate PDE. (f) Manage access to the stored PDE. Each of the listed components of the ONW system are the various aspects of the functional requirements specifications that define the core of the ONW system that must be identified at the ONW system's design phase.

5.1.1 The citizen's interface

The citizen's user interface actor of the ONW system is mapped to the use case to capture PDE, upload PDE and receive acknowledgements within the ONW system as depicted in Figure 6. The citizen is the person in the street who captures and uploads the PDE of an incident, and is referred to as the uploader. The uploaders capture PDE of crime using their mobile device and upload the captured PDE to the ONW repository. The uploaders receive an acknowledgment at successful upload of PDE, when their PDE is downloaded to be used as potential evidence in neighbourhood crime investigation, or when PDE is presented as evidence in a court of law.

5.1.2 The law enforcement agents interface

The function of the law enforcement agents is to download PDE to corroborate their physical scene investigation and/or validate the evidence. The law enforcement agents must manage the PDE by either accepting the evidence as valid and useful to their investigation, or by rejecting and discarding PDE when inconsistency is discovered in any evidence (Omeleze and Venter, 2015). As depicted in Figure 6, the role of the actor as law enforcement agent is bound to that of the digital forensic investigator who may assume both roles when necessary or based on job description.

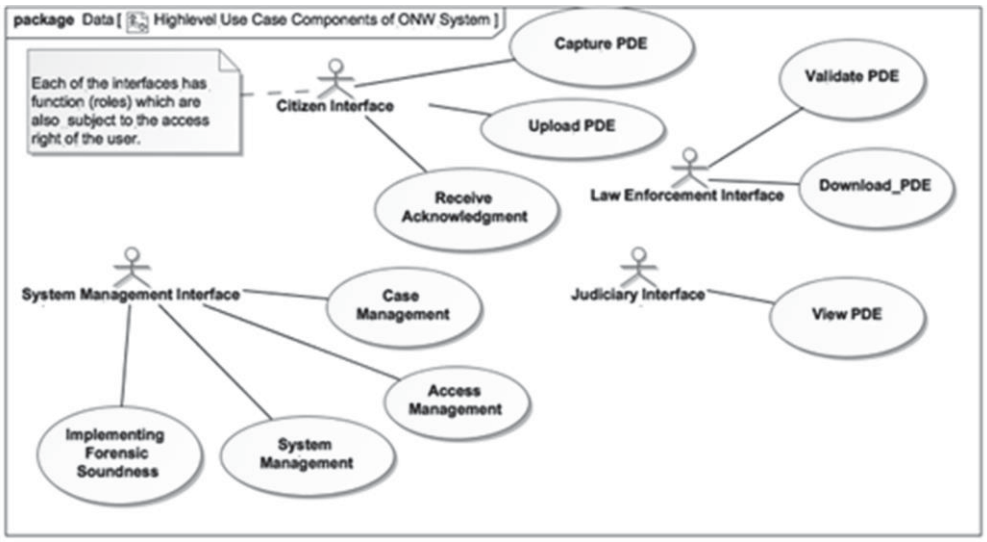


Figure 6: ONW system's users interfaces and their functions

5.1.3 The judiciary interface

The judiciary members function can be carried out by the court clerk or other authorised person. In practice, a situation may arise where PDE captured and stored is to be utilised as real evidence in order to shed light on a case in a court of law. The presiding judge, court clerk or legal counsel of the plain ff or defendant may need to view the alleged PDE, which can be made available to all parties involved. The role of the justice system is a requirement for the ONW system to successfully operate within the legal constraints as it concerns using PDE stored in the ONW repository. Further- more, when PDE from the ONW repository are made available, it assists both the prosecutor's and the defendant's legal teams to prepare legal arguments. Such PDE can be one of the avenues to enhance the understanding of a case to be adjudicated upon by presiding Judge, as to what happened during an alleged neighbourhood crime.

5.1.4 System management interface

The ONW system ensures that acquired PDE upholds the information security services mechanisms that encompass confidentiality, integrity, authorisation, authentication and non-repudiation, and the properties of forensic soundness indicators (FSI) at all levels of the ONW system's components interaction. The purpose of the system management interface function is to a ain: (i) Confidentiality - Restricting PDE usage to the intended and authorised audience (i.e. the law enforcement agents, the digital forensic investigators and the judiciary) (ii) Integrity - Ensuring that PDE is not altered in transit between the uploader and the downloader or during storage. (iii) Non-repudiation - Upholding the PDE uploader's original intention. For example, the uploader of a PDE may not deny at a later me of his or her intentions in the creation or transmission of the PDE. (iv) Authentication - The digital data origin, and the identity of the uploader and receiver are confirmable. The functional requirements features are addressed using the architectural requirements of the ONW system which are discussed next.

5.2 Architectural requirements of the ONW system

The architectural requirements are the quality requirements, architectural patterns and architectural strategies. These are used to realise the identified functional requirements. Based on the functional requirements, the

quality requirements for the ONW system are: a) Security requirements b) Reliability requirements c) Usability requirements d) Auditability requirements.

5.2.1 Security requirements

The security requirements of the ONW system are achieved in conjunction with the information security services mechanisms i.e., confidentiality, integrity, authentication, authorisation and non-repudiation (CIAAN) (Susanto et al., 2011). The measures employed to implement CIAAN are: (i) Confidentiality - which is realised using encryption. (ii) Integrity is realised using cryptographic hash function. (iii) Authentication is realised using session authentication, username and password. (iv) Non-repudiation is achieved using digital signature. To ensure the security requirements of the ONW sys- tem are addressed, layered architectural patterns are used at the high level of the ONW system as depicted in Figure 7. Furthermore, microkernel and pipes and filters architectural patterns are employed at layer access, business and persistence layers of the ONW system (see Figure 7). Using these architectural patterns in conjunction with architectural strategies like cryptographic hash, digital signature, encryption, decryption and port lockdown, limiting of access and exposure inherently addresses the security requirements of the ONW system.

5.2.2 Reliability requirements

The architectural strategies to achieve reliability requirements are: (i) Firstly, fault prevention - this is by adhering to thorough system testing and the use of resource locking as well as the removal of single points of failure (Len Bass, 2012). (ii) Secondly, faults are detected using deadlock detection, logging, checkpoint evaluation and error communication architectural strategies. (iii) Clustering and throttling architectural strategies are also used to avoid resource bottlenecks which are created due to components overconsumption of resources. (iv) Meanwhile, a microkernel architectural pat- tern is used to achieve reliability at the second level of granularity with the microkernel's bus. The microkernel's bus maintains reliable communication channels between all access points, detects alteration of PDE at upload and maintains a tamper-proof system. (v) Pipes and filters architectural pattern is used to achieve reliability with the pipes encapsulating the encryption and decryption features of the system and at the same managing attacks such as eavesdropping and data interception.

5.2.3 Usability

Usability measures the effort a user requires to use the ONW system. The ONW system provides users with an easy to use and familiar system, which anticipates users' needs. For example, the ONW system makes it easy for its users to perform tasks on the system such as remembering the last typed URL and recalling usernames and passwords (especially when using the same device). Some of the architectural strategies to realise usability include separation of concerns which is embodied in the model view controller (MVC) architectural patterns and design patterns, such as flyweight pattern. These usability functions enable the design of the ONW system's user interface that allows for a component-oriented design, thereby restricting dependence, to avoid components re-designing when the need arises to change the user interface.

5.2.4 Auditability/monitorability

Auditability provide services that enables the ONW system to log the input and output ac vi es of the internal and external resources. This is in order to track system's and user's activities. The auditability functions of the ONW system are in place to maintain chain of custody, as to who did what and at what me within the system. Therefore in the unlikely event of system failure/crash, the ONW system is able to roll back to its last stable state, using its audit log information. Information, such as location, date, time stamps used to re-trace the ONW system's activities.

5.3 Architectural patterns used for the structural design of the ONW system

Using the various identified architectural patterns and strategies that best address the quality requirements of the ONW system, an architectural structure is created as shown in Figure 7. The ONW system architectural structure is based on a three- tier layered architectural pattern. Each of the layers is enclosed with other identified architectural patterns to address the quality requirements of the ONW system.

Figure 7 shows a layered architecture, the parts labelled B & D relies on pipes and filters pattern to ensure secure messaging channel, encryption and decryption processes while microkernel architectural pattern (labelled C) is

also used to sup- port the security features of the ONW system. The microkernel and pipes and filters architecture are encapsulated in the layered architectural pattern to concretely address security, reliability and auditability quality requirements. For example, for the encryption of PDE at upload and the decryption at download, the pipe (i.e. labeled B) is used as the messaging channel. The microkernel bus (label C) is the business logic of the system that provides stability functions, as well as the adapter to allow for connection of additional components between the access layer and the persistence layer.

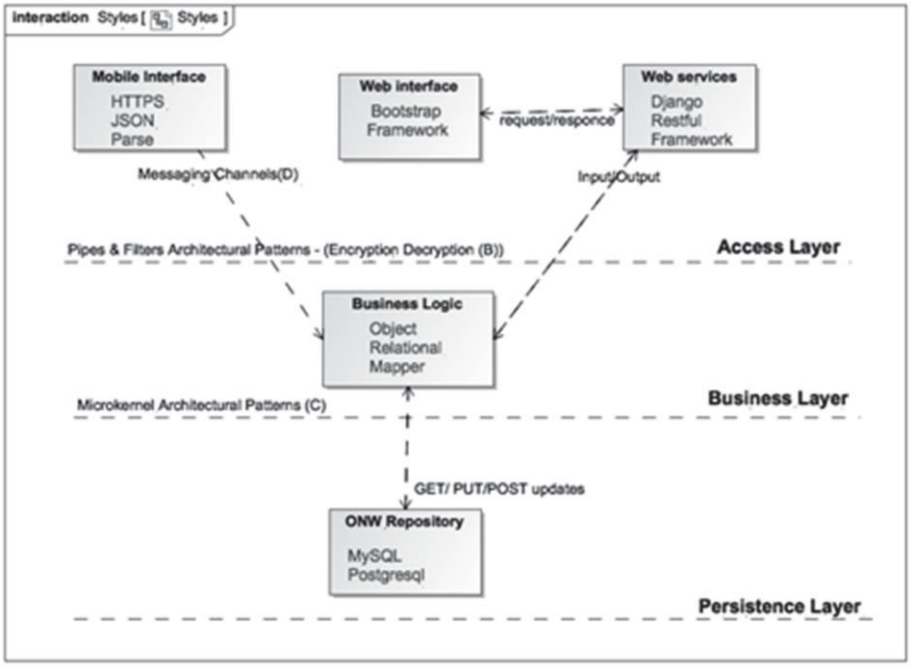


Figure 8: High level architectural design of the ONW system

One of the advantages of using layered architectural pattern is to realise flexibility and pluggability. This allows for a technology neutral design, then freeing the system from technology lockdown (Solms and Loubser, 2010). Other merits is the ease of adding components at the access layer without the need for a architectural re-design. As depicted in Figure 7, the web front-end of the ONW system is accessible via mobile or desktop clients, which are designed using MVC architectural style. The persistence layer holds the data access layer that encapsulates and exposes data for access by other layers of the architecture. It provides for the application programming interface (API) that presents a means to manage the stored PDE in the ONW repository. The next section evaluates the ARES process in light of the case study – ONW system.

6. Evaluation

The ARES process adds a layer of abstraction to accommodate device upgrade instead of building a new application. Therefore, in evaluating the ARES process, the basic factor considered is its impact in designing the ONW system. The ARES process narrows the focus of the ONW system to the essential aspects of the system in order to design a forensically sound application by identifying its core requirements, aligning it to the stakeholders needs, as well as addressing the system's constraints.

By adopting the component-based design approach, the ARES process focuses on harnessing advantages such as the following:

(i) The ARES process makes it possible to accommodate the various challenges that arise due to changes in legal standards and legislation since it focuses on architectural design at the commencement of a DF application design, as well as in the process addressing the jurisdictional or multi-tenancy constraints surrounding digital evidence acquisition and storage.

(ii) The ARES process enables the accessibility of DF application design, thereby allowing for easy evaluation of the DF application's techniques by peers, the justice system and other stakeholders to ascertain the forensic soundness of the DF application. Previous studies have shown that software failure originates mostly from inadequate requirement engineering specifications (Hofmann and Lehner, 2001). This implies that failure could

be averted, or reduced to a minimum level by employing the proposed ARES process in designing DF applications like the ONW system.

(iii) The ARES process exposes the ONW system's core to all stakeholders at the design phase to facilitate decisions that align all constraints to the system's requirements. This ensures an effective and extensible design of digital forensic applications, where the quality of the final product is measurable with respect to the system's expected output. Such measurements are defined by concretely quantifying the capabilities of the DF application's requirements.

The shortcoming of the ARES process is defining and adding some South African legal requirements constraints to the design of a DF application. For example, the Privacy of Personal Information(POPI) Act, Act 4 of 2013, requires that, where citizen's information is acquired to avert crime or for national security circumstances, adequate measures must be employed to protect this information (POPI-Act, 2013). Yet these measures are not clearly defined in the PDE capturing process of the ONW system.

7. Conclusion

DF applications are used to reach critical decisions during criminal investigations, and therefore their effectiveness and accuracy must stand up to scrutiny inside and outside of court. To ensure effectiveness and accuracy of DF applications, there is need for employing architectural requirements engineering specifications in their development. This paper has proposed an architectural requirements engineering specifications (ARES) process for developing DF applications.

The ARES process can lead to a standardisation of DF application design. This leads to greater user confidence in DF applications and their credibility is less likely to be challenged in court. The ARES process recognises the key role that users of any DF application play and so the process begins with identifying the users' needs. These needs are aligned to the quality requirements of the application, while taking cognisance of any constraints that must be considered. Overcoming the challenge of the constant upgrading of devices and technological changes is a key element of the ARES process. It highlights the processes a DF application undergoes to fully adapt to pluggability, extendibility, maintainability and modifiability. A case study to apply the ARES process is implemented using the process to design the ONW system, which has been developed to tackle neighbourhood crime in South Africa. Applying the ARES process in designing the ONW system, has shown an easy to follow process for designing DF applications, and thus makes an effective contribution to the field of digital forensics.

Acknowledgements

This work is based on research supported wholly/in part by the National Research Foundation of South Africa (Grant Numbers 88211, 89143 and TP13081227420).

References

Eoghan Casey. Digital Evidence and computer crime Forensics science computers and the internet. Elsevier Inc, third edition, 2011. ISBN 9780123742681.

Fred .A Cohen. Digital Forensic Evidence Examination. Fred Cohen and Associates out of Livermore, third edition, 2009. ISBN 9781878109446.

Simson Garfinkel, Paul Farrell, Vassil Roussev, and George Dinolt. Bringing science to digital forensics with standardized forensic corpora. digital investigation, 6:S2--S11, 2009.

Government-Gaze e ECT-Act. Electronic Communica ons and Transac ons Act, Act 25 of 2002. Technical report, PDF Scanned by Sabinet [Online - Accessed 08 February, 2014], August 2002. South Africa Government Gazette - Legislation - South Africa - National/Acts and Regulations/E/Electronic Communications And Transactions Act No. 25 Of 2002/The Act.

Government-Gaze e POPI-Act. Privacy and data protection - discussion paper 109 (project 124) - south african law reform commission (2005-10). Technical report, [Online - Accessed 08 August, 2014], August2013.URLhttp://www.sabinetlaw.co.za/justice-and-constitution/legislation/ protection-personal-information.South Africa Government Gazette - Legislation- South Africa - National/Acts - Privacy and data protection Act No.4 of 2013.

Daniela E Herlea, Catholijn M Jonker, Jan Treur, and Niek JE Wijngaards. Integration of behavioural requirements specification within knowledge engineering. In Knowledge Acquisition, Modeling and Management, pages 173--190. Springer, 1999.

Hubert F Hofmann and Franz Lehner. Requirements engineering as a success factor in software projects. IEEE software, 18(4):58--66, 2001.

Dean Leffingwell and Don Widrig. Managing software requirements: a use case approach. 2003. ISBN-13: 978-0321122476, ISBN-10: 032112247X.

Rick Kazman Len Bass, Paul Clements. Software Architecture in Practice, 3rd Edition. Addison-Wesley Professional. Part of the SEI Series in Software Engineering series, 2012. ISBN-13: 000-0321815734 ISBN-10: 0321815734 3rd Edition.

Stacey Omeleze and Hein. S Venter. Testing the harmonised digital investigation process model using an android mobile phone. In Information Security for South Africa, 2013, pages 1--8. IEEE, 2013.

Stacey Omeleze and Hein. S Venter. Towards a model for acquiring digital evidence using mobile devices. In Tenth International Network Conference (INC 2014) and WDFIA 2014 Plymouth University, UK, pages 1--14. Plymouth University, UK, 2014.

Stacey Omeleze and S. Hein Venter. A model for access management of potential digital evidence. In 10th International Conference on Cyber Warfare & Security (ICCWS), pages 491--501. CSIR, University of Vender and Academic Conferences Limited, 2015.

Suzanne Orofino. Daubert v. merrell dow pharmaceuticals, inc. the battle over admissibility standards for scientific evidence in court. J. Undergrad. Sci. 3: 109-111(Summer 1996), 1996.

Klaus Pohl. Requirements engineering: fundamentals, principles, and techniques. Springer Publishing Company Incorporated, 2010. ISBN:3642125778 9783642125775.

Pamela-Jane Schwikkard and Steph E Van der Merwe. Principles of evidence. Juta and Company Ltd, 2009. ISBN: 978 0 7021 79501.

Fritz Solms. What is software architecture? In Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, pages 363--373. ACM, 2012.

Fritz Solms and Dawid Loubser. Urdad as a semi-formal approach to analysis and design. Innovations in Systems and Software Engineering, 6(1-2), 2010.

Heru Susanto, Mohammad Nabil Almunawar, and Yong Chee Tuan. Information security management system standards: A comparative study of the big five. 2011.

Aleksandar Valjarevic and Hein S Venter. Harmonised digital forensic investigation process model. In ISSA, pages 1--10. IEEE, 2012. URL DBLP:conf/ISSA/2012.

Murdoch Watney. Admissibility of electronic evidence in criminal proceedings an outline of the south african legal position. Journal of Information, 2009 .

Non Academic Papers

Protecting Real-time Transactional Applications With DDoS Resistant Objects

Hugh Harney[1] and Robert Simon[2]
[1]Axiom Inc., Columbia, USA
[2]Department of Computer Science, George Mason University, USA
hh@axiom-inc.com
simon@gmu.edu

Abstract: DDoS Resistant Objects (DRO) are a distributed information sharing infrastructure that is tolerant of disrupted communication environments and resistant to DDoS attacks. DRO uses advances in data security to enable distributed data management techniques like: local secure caching of data for retrieval and search, localized network reconstitution in the face of successful DDoS or cyber attacks, and advanced data search and request capabilities for advanced user capabilities. Today existing systems like Internet Relay Chat (IRC) provide multicast-like service using hub and spoke architectures and multiple, frequently unprotected, pairwise connections. This results in a network that provides DDoS attack vectors via network sniffing, message retransmission, and control message spoofing. Furthermore, due to security considerations all retransmissions must originate from the original data source. This makes reconstitution of operational networks highly inefficient in terms of system bandwidth and management of pairwise interconnections. DRO integrates emerging technologies in a synergistic manner to create capabilities that exceed the current distributed data management systems. DRO has 5 areas of technical innovation. DDoS Resistant Multicast (DRM) automatically identifies and filters messages that are out of sync with network management. It binds the multicast address to a group cryptographic key enables the address to inherit several desirable characteristics of the underlying key – controlled frequent updates, rollover capabilities, group topologic manipulation, group membership control, secure command and control, and cryptographic level entropy. Delay Tolerant Multicast allows operation during attacks. Using the object structure of the DTN Bundle Security Protocol with the DRO data distribution architecture facilitates delay tolerant multicast data delivery without the overhead of the DTN communication protocols. Multiple partition secure objects allow the network to securely cache, search and retransmit DRO objects. The object security key management is divorced from session security enabling long lifetime object structures that embrace attribute based access control key management that permits late binding of destination and late admittance to access groups. Distributed caching and retransmission allows reconstitution of network operation using localized resources. The DRO architecture incorporates caching and retransmission service at network and application layers. This facilitates network reconstruction and resiliency using local resources. Data searches are enabled by the encrypted metadata payloads enabling, intra-object referencing and management. Decentralized data searching that preserves security of the critical information provides a new paradigm for secure management of mobile information.

Keywords: distributed denial of service attacks, cryptographically protected networking, secure multicasting, key management

1. Introduction

Distributed Denial-of-Service (DDoS) Resistant Objects (DROs) and the multicast delivery architecture that supports them represent a radical re-examination of how information is managed for real-time transactional services. The DRO system operates both in benign and DDoS stressed environments. DROs facilitate a data object delivery architecture that provides robust and secure object delivery in the presence of DDoS. These objects implement group secured data and Metadata partitions to facilitate store and forwarding service from decentralized caching nodes in the architecture. The result is a flexible, secure and resilient architecture for data delivery.

Current real-time transactional systems assume robust communications and adversarial challenges that do not include cyber sniffing or cyber attacks. As a result, many current systems, such as Internet Relay Chat (Lu 2008) do not provide communication allowances for delayed or disrupted communications or servers. Any major disruption results in widespread network-level infrastructure peering relationship re-establishment and user-level retransmission of disrupted data. This opens a tremendous amplifier for DDoS attackers.

DROs frustraste a DDoS attackers' attempts to disrupt the multicast network. It makes the organization of a flooding attack more difficult and provides a natural process for filtering messages out of the network. If we assume an adversary finds some avenue to disrupt communication services, DRO provides a Delay Tolerant Enhanced transport layer routing. Delay Tolerant routing takes mechanisms developed within in the Delay Tolerant Networking research community and adapts them for multicast message retransmissions. The multicast

structure facilitates lower overall network bandwidth load, multi-path routing, and real-time routing management, which greatly narrow an attacker's window of opportunity.

DRO recovery mechanisms are distributed and provide a degree of local reconstitution in the event that an attack disrupts a DRO server. This reduces the system-level impact of such events. This local resource capability allows reconstitution of user operations for multiple missions, without the self-flooding common in today's systems. Local reconstitution is enabled by the multiple-partitioned, protected data structures -- the DROs --, which carry information through the multicast overlay network. Common transactional services are transferred to the DRO architecture by moving message sets to the mission data payload within the DRO.

In the delay tolerant multicasting scheme presented here, the multicast addresses are bound to grouped key. This grouped key provides address agility, address filtering of "out of sync" messages, unlimited address updates, virtually zero predictability for adversarial action, and an asymmetric advantage over the adversary due to the limited attack time window. Grouped keys facilitate attribute-based cryptographic key access controls, which allow for late binding of destination (post message transmission) and enable local retransmission of data, with no loss to data security.

DRO uses distributed secure repositories to facilitate discovery of useful information to users. This information can be presented in a subscription manner similar to popular mobile applications. In addition, the use of multiple encrypted partitions in object design reduces the systematic information disclosure risk if any cache is compromised. Only the Meta data concerning any object needs to be available to the caching and retransmission components, the actual information is only reveled to the end user. The user capabilities enabled by DRO will bring "Twitter-like" information tagging, searching, chaining, poly references, stream creation, stream delivery, and stream customization to data that is sensitive, with no loss of mission data protection. Additionally, multiple complex inter-object linkages and connections can be expressed in the DRO system data to provide users with rich context for information cognition and management.

2. Operational concepts

DRO is centered on a data structure that enables a host of DDoS resistant capabilities when combined with other emerging technologies. To illustrate how this architecture interacts refer to Figure 1, High Level DRO Architecture.

2.1 Initiation

Prior to messages being passed, a Transaction Authority (TA) establishes the parameters for and characterizes transactions authorized to take place in DRO. These parameters include the key management policy, user attribute requirements, priority, and resource allocation to support recovery.

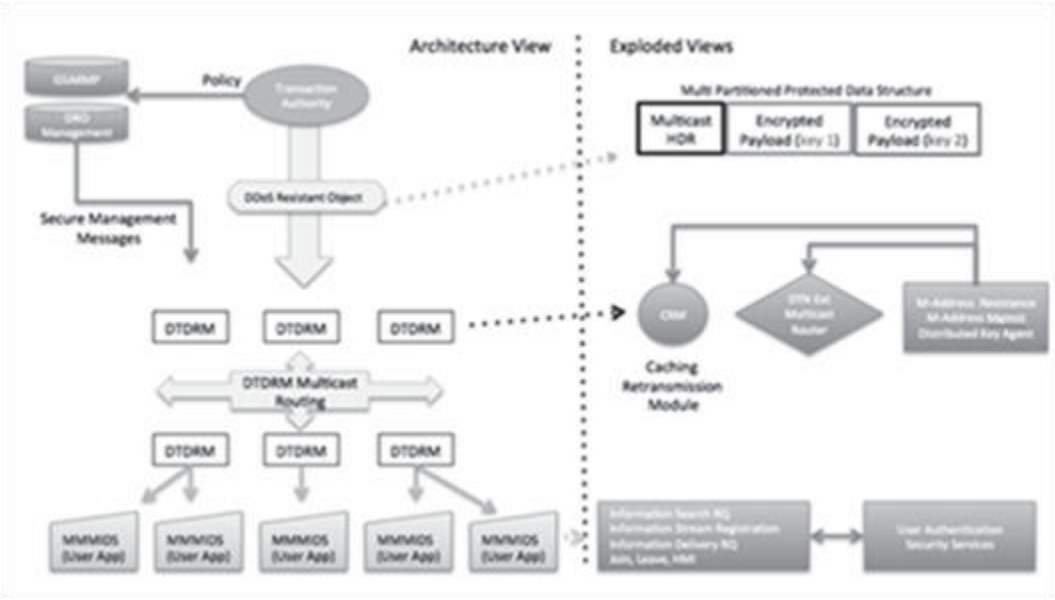


Figure 1: High level DRO architecture

The security and information policies are sent to the Group Secure Association Key Management Protocol (GSAKMP) (Harney 2008) and DRO Management elements. Using the policy parameters, GSAKMP generates appropriate keying materials and through DRO management creates DRO multi-partitioned protected data structures (DRO objects) to send the management information to the Delay Tolerant DDoS Resistant Multicast Routers (DRM). Once the DRM processes the management messages, all policies and keys will be in place to operate.

2.2 Operation

Users interact with an authorized Mobile Multi Media Information Distribution System (MMMIDS) application. The MMMIDS applications can present a legacy human interface for applications like IRC and then transmit application-specific transactions in DRO objects. However, MMMIDS can also facilitate modern distributed information paradigms (similar to Twitter) where the user is presented with streams of information based on their personal preferences. Each stream is comprised of metadata payload information. This stream could include almost anything that will help users understand and manage the data available to them, such as human readable descriptions, tags, or links.

In operation, a data source, or user, will create data that is transformed into a DRO object. For illustration purposes, the creating entity in this case will also be the TA, but can be any user or system node. This DRO Object is a data structure that uses 2 grouped keys:

- The *group key* protects the metadata payload, which provides searchable mission data payload information. The metadata is at a lower common security classification then the content data
- The *mission data key* provides actual consumers with access to the mission data payload

Once the object is created, it is sent to the appropriate multicast address. The DRM routers will transmit the message provided the key-derived address is within a rollover window. Each DRM also has the capability to cache and retransmit DRO objects through the Cache and Retransmit Module (CRM). The management agent identifies the multicast group DROs that a CRM module is responsible for caching, the time interval for the cache, and the subset of the group to which the CRM is responsible for retransmitting the objects.

After the DRM routers transmit the DRO object, two methods of receipt of the object are available. As a first option, the user application (MMIDS) can receive the object directly because the application was listening to the multicast channel. In the second option, the user can cause the object to be retrieved as part of a search/retransmit request. In either case, an authorized user may log onto the MMMIDS and request information access. The application can create complex command requests for searches across the CRM Metadata store.

Once the user selects an object for retrieval, the appropriate secure command is issued. The CRM retransmits the selected cached object on an appropriate address (pairwise or multicast). The user's credentials are used to retrieve the mission data decryption key. The mission data payload is decrypted, as close to the user as possible, to minimize data exposure.

While this scenario shows one object on a single multicast address, the operational system supports many objects and many multicast addresses. In the operational system, a single MMMIDS user can also be both a Transaction Authority on one address, and a simple receiver on another.

3. DRO architecture

Object-based communication protocols have been shown to operate in difficult network environments, including deep space and tactical radio networks (Koponen 2008). The IETF Delay Tolerant Networking (DTN) research group explored this for pairwise communications. [REF] DRO intends to expand this research to address resistance to DDoS in the Multicast protocol suite.

A DTN Enhanced DDoS Resistant Object is an atomic data object made up of two parts – protected metadata and protected mission data. The inclusion of multiple cryptographic protections in the base DRO acknowledges the DoD's requirement for secure data systems and enables a wider range of network services, including single level object searches, widespread single level object caching, and distributed object retransmission.

The components of a DRO are:

- IPV6 Multicast address: To outside observers, this appears to be a standard address. Its properties will be discussed in the DDoS resistance section.
- Metadata payload header: This provides the information needed to decrypt the Metadata payload.
- Metadata Payload: Metadata is searchable data about the mission data payload
- Mission Data Header: This provides information needed to decrypt the Mission Data Payload.
- Mission Data Payload: This is data for consumption by users. This data is frequently much more sensitive than the metadata

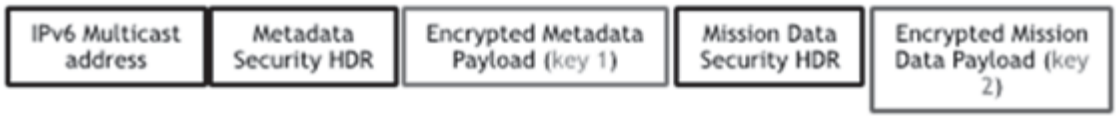


Figure 2: DRO multi-partitioned protected data structure

Encryption properties of Metadata and Mission Data payloads impact the delay tolerance of the communications. Group cryptographic keys are used for both the metadata and the mission data payloads. This allows multiple recipients to process a single message. The keys for each payload are different, facilitating different access control policies for metadata vs. mission data.

3.1 Grouped cryptographic keys

Grouped cryptographic keys are a well specified, but under-utilized, technology. The technology has been standardized in a number of RFC, including RFC 2092 “Group Key Management Protocol (GKMP) Specification”, RFC 2094 “Group Key Management Protocol (GKMP) Architecture”, RFC 4535 “GSAKMP: Group Secure Association Key Management Protocol”, RFC 4524 “Group Security Policy Token v1”

Grouped cryptographic keys have the following advantages of providing a distributed peer based architecture, a delegated policy based management, and Attribute Base Access Control (ABAC). ABAC allows the key management system to grant keys after the message has been transmitted. Access control is based on verifiable attributes of the person or system receiving the key. These features allow the system to cache and deliver messages to authorized recipients and provide a secure cryptographic management policy.

Another aspect of this approach is trust management. Management of trust in a group is always a focus. Logical Key Hierarchies (LKH) provide an organized cryptographic key structure to exclude untrusted group members and reconstitute a trusted group or groups using a small number of multicast messages. LKH is adept at changing the topology of groups. This topology manipulation has useful properties in responding to a cyber attack

The management of group cryptographic keys implies some desirable properties for delay tolerance. Group keys permit attribute-based access control policies, thus allowing the binding of destinations after the transmission of an object. This binding facilitates multicast network retransmission of cached objects to reduce the impact of network and endpoint re-synchronization after, and during, communication outages.

4. Delay tolerant DDoS resistant multicast (DRM)

DRM assumes the communications network will experience cyber and DDoS attacks. DRM mitigates this bandwidth uncertainty by incorporating features found in delay tolerant pairwise networking protocols. DRM bases the core data transmission protocols around data objects to facilitate a data delivery paradigm, which passes object delivery responsibility to a network using a hop-by-hop approach. This hop-by-hop approach eliminates the need to maintain a long-term high bandwidth channel from source to destination. Each of these hops individually implements a best effort object delivery. Based on the current network environment, this facilitates optimization of real-time object delivery.

Further, DRM modifies traditional multicast to make it DDoS Resistant. The modifications are the multicast address a function of group keys, provide key dependent pruning of multicast traffic policy, offer integration of group keying, and simplify the management of group keys.

A critical modification to multicast is the inclusion of a variable bound to a group key into the IPV6 multicast address. This makes a portion of the multicast address difficult to predict for someone without knowledge of the particular group key. Additionally, the multicast address inherits several management properties of group keys. The address changes are provably secure and distributed and can change with the frequency of a group key. The group key tied to the multicast address can be updated numerous times, thus facilitating a change in the multicast address.

The DRO system provides a message pruning system that significantly limits the opportunity of an adversary to perform denial of service on a group address discovered by an attacker. An attacker may attempt to use that address to either replay legitimate, older messages over the address or perhaps flood it with other traffic. Similar to key rollover policy, this pruning mechanism is based on key, time, and key versions. It allows routers to route messages over the current group address, *N*, (which changes periodically as a function of time) or over a set of previous *c* addresses, where *c* is a configurable parameter. Should a DRO come in for the group, a multicast message would be forwarded using keys N, (N-1), (N-2) … (N-c). The window of opportunity for an adversary to use a discovered address would quickly be aged out of the allowable set of addresses. The tunable parameter, c, can be optimized for network performance or environmental factors and is a subject of research under this proposal.

Management of group keys also manipulates the topology of the groups. Group key management tree construction (e.g., LKH) allows a single message to re-key an entire group of keys. In fact, the group topology can be manipulated with very few control messages; a single group can be split into many groups, or many groups can be combined into one single group. This implies that the multicast address routing can be manipulated in a similar manner to provide distribution or multi-path delivery of data in times of system stress.

4.1 Justification and rational

Why Multicast? The use of DRM allows a single message to be distributed to all the desired destinations. This implies a promiscuous-mode listening delivery scheme that obfuscates any role analysis requisite for a DDoS attacker to determine mission critical nodes and links. Multi path same message delivery makes DRM less vulnerable to single node or single link attacks. Based on the group key topology management, the agility of the multi-path delivery trees makes the DDoS attack more difficult to plan or persist in the presence of automated active group key management.

DDoS attack architectures uses a botnet to flood or attack a network. This requires that all attacking hosts on the botnet construct messages usurping the network resources such as, bandwidth or processing power. DRM makes it difficult for the DDoS attackers to construct messages that would usurp the network's resources. The construction of system routed multicast messages implies the attacker must get access to a current address and distribute it to the botnet in a time window that will allow the attack to progress. DRM frustrates this in several ways:

- Any multicast message with an address outside the allowed range will be pruned
- Group key management protocols are secure and difficult to gain information from using standard techniques
- The shelf life of any copied key can be made arbitrarily short by the judicious selection of the variable *c*, in the address roll over scheme
- Group key management re-key technique can change the group address key in a very short time if a compromise is detected
- Time To Live is replaced with address agility. This is difficult to predict or spoof.

DDoS attacks against a network also target the network management and control. This implies that any centralized single point of failure would be a prime target to an attack. To eliminate one possible attack vector, many of the address agile architectures in the academic literature (e.g., OpenFlow Random Host Mutation, Jafarian 2012) rely on a single point of address distribution. DRO uses a distributed key management framework with distributed computation of addresses. Another attack vector for network layer protocols is exploitation of control channel functions. Our approach mitigates this type of threat because a DRM control messages are encrypted as objects reducing the probability that they can be malicious.

4.2 Caching

One major advantage in the DRO system is the ability to cache objects and subsequently search them. DRO contain a metadata payload that describes the mission data payload in two respects. First, the mission data payload is described in network terms to facilitate network layer retransmission. Second, the mission payload is described in non-sensitive application layer terms to facilitate application layer caching.

Network layer caching addresses a major DDoS issue with requiring retransmissions to originate from the original source. Using the Metadata field in the DRO the network can construct caching resources throughout the network. This is a manageable capability and multiple caching architectures can be conceived to facilitate many different operational requirements. These caching resources could also be constructed in an architecture that is hierarchical to support enterprise reconstitution and management.

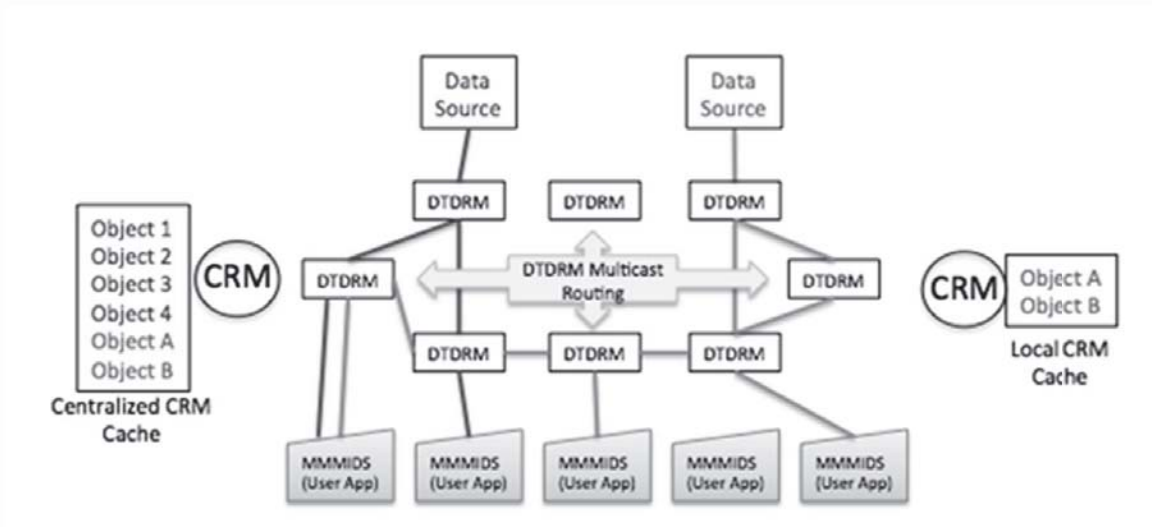


Figure 3: Hierarchical CRM resources with tailored caches

The goal of creating caching/retransmission resources is to minimize the latency and bandwidth requirements for serving the user retransmission requests. This approach is particularly applicable if applied to users who need to join a group and "get caught up" with the recent object traffic for that address.

This caching and retransmission method requires a re-design of the underlying retransmission request function in most transport layer protocols. This function treats retransmission in a manner similar to service discovery. The host requiring an objects retransmission sends out a control message to a control multicast address. This request can operate in a planned or pure discovery manner. In planned architecture the user has exact knowledge of the retransmission point (i.e. address and hop distance) and the user constructs and sends a retransmission request. The retransmission point either has, or collects, the objects required and transmits them to the user. This retransmission can be multicast or unicast.

4.3 Caches and searching as advanced computing elements

The DRO system architecture creates a unique capability for data sharing, management, search, storage, and retransmission; namely, in the cache/retransmission module (CRM). The CRM uses the object paradigm and creates a mechanism, for object sharing between computers, which is secure, timely, and managed using modern metadata.

The CRM can scale to meet a wide variety of retransmission scenarios. CRM is a cache of data objects, the metadata key, and an ability to search unencrypted metadata for an object. It is made up of basic DRM capabilities and to some extent exists at any DRM router. This simple concept can be scaled up to provide an object library that caches a vast number of objects. Alternatively, it can be scaled down to provide a minimal cache of multicast traffic for local retransmission.

CRM facilitates local data object retransmission. The DRM approach embeds messages into the mission data payload. This enables CRM to cache and retransmit, as the system requires. This configuration solves many DDoS vulnerabilities in the current IRC structures.

4.4 Mobile multi-media information distribution systems (MMMIDS)

Popular commercial information sharing systems offer subscribers "streams" of information they can follow. These streams offer general descriptions of more detailed information objects that the user can access. The DRM architect creates a direct support system for MMMIDS. The CRM elements of DRM provide a capability to cache and search objects for delivery. The CRM in this case goes beyond commercial applications in that it also provides security services facilitating search at a single "common low" level of classification, while allowing the delivery of protected mission data at various categories of classification.

The CRM will make use of multicast addressing to facilitate distribution of stream information. The users can listen to a particular stream ID on the multicast channel. Once the user reviews the stream data related to their choices, they can select to retrieve an object and the CRM simply delivers the DRO message payload. If the mission payload is encrypted (encryption can be null for non-classified payloads) the user will have access to the decryption key and, if done properly, the decryption will take place as close to the authorized user as possible.

4.5 DDoS multicast peer to peer network routing

One of the major challenges in designing Delay Tolerant DDoS Resistant Multicast (DRM) is to make it highly resilient to extended DDoS attacks while maintaining high performance during periods of normal operation. Although both layer 3 and overlay multicast has been heavily studied, relatively little attention has been paid to understanding the performance and engineering tradeoffs of multicast protocols and their implementation support under conditions of repeated and extreme denial-of-service (Sterbenz 2013). This is especially true for DTN multicast (Uddin 2011).

Once an object has been cryptographically validated by demonstrating possession of a valid multicast address, a DRM router needs to make a forwarding decision. In particular, at any one time, a router will have several possible neighbors on its overlay tree to forward the object, including a neighbor(s) on a shortest path unicast path or a multicast tree. However, by assumption, some of these overlay links and neighbors will be subject to cyber attacks. In probabilistic forwarding, a node attaches some probability as to whether to forward an object. This is a variant of a well-known technique called epidemic routing (Ramanathan 2007.) Epidemic routing widely used as a flooding method in the context of intermittently connected networks such as DTNs. Probabilistic forwarding is best suited to the scale of networks envisioned for DRM.

One way to improve delivery performance is to use multiple copies of the same object within the network [Xue09]. Each copy can take a different path along the overlay, thereby increasing the likelihood of delivery as well as decreasing the delay. One common control technique is to limit the total number of copies using counter-based mechanisms. DRM adopts this practice. This approach facilitates object storage in multiple searchable caches consistent with the searchable nature of DRO object availability

Pure peer-to-peer overlay systems are known to be difficult to manage and to effectively scale. Scalability is generally achievable via the introduction of a hierarchy. For this reason DRM will support the use of specifically designated "core" nodes (Abdulla 2007). Such a node has special responsibilities within the DTN object-forwarding paradigm. Core nodes simplify DRO network management functions such as multicasting, group membership and network security, by placing these activities inside of cores. For instance, non-core nodes can manage their DRO security associations via interactions with core nodes, rather than relying on pure peer-to-peer approaches. Further, core nodes can be tasked with the responsibility for CRM actions such as querying and retransmission. Note that cores are *not* single points of failure, as the functioning of the cores themselves is fully distributed.

We consider a variety of DRM object forwarding policies for core nodes. One type of policy called Direct Forwarding Plus Core, or DFPC for short. In this policy, in addition to each node performing probabilistic and multi-copy forwarding, it will (perhaps probabilistically) forward objects to a core node. A different forwarding

policy is full or partial core-to-core, whereby two cores exchange some or all of their stored data. As it receives new objects cores will be re-forwarding decisions to unicast or multicast groups.

5. DRM analysis

We now provide an analysis of the effectiveness of the DRM multicast address rollover mechanism in preventing DDoS attacks. We assume that the attacker is a botnet under the control of a botmaster. Attacks will be automatically launched once the botmaster obtains the current multicast address used by DRM. Upon receiving this information the botmaster immediately attempts to contact all bots with the new address, and orders them to attack using this address. As noted in Section 4, the length of time that an address is valid for, as well as the parameter *c,* the number of valid addresses for which nodes can accept new messages, is a critical tuning knob. Reducing the length of time *T* the address is valid for, and minimizing the value of *c*, will reduce the effectiveness of the DDoS attack. However, excessively small values for either *T* or *c* will reduce the number of messages received by the system, due to network conditions such as congestion and packet delivery delay.

We present an analysis to understand these trade-offs. Since the IPv6 specification allows group ids to be up to 112 bits long, we assume the probability that the attacker can predict the next address or that the next address will collide with any existing valid addresses is effectively zero. Assume that the botmaster instantly is able to obtain a new DRM multicast address. It then must contact all of the bots with this information, and the bots must launch the attack. The probability of success is a function of whether the bots are awake, and whether host and network conditions allow for reception of the address before it slides out of the window. Further, if the window interval for a valid key, or the value of *c* is too small, then network conditions will decrease the chance of DRM nodes receiving a valid message in time. To numerically analyse these probabilities of network conditions and bot state we can use a two-parameter gamma distribution. By changing the values of these parameters a wide variety of network and bot conditions can be modelled. An example of this approach is now described. Assume that there are 2000 DRM nodes, and on average each node generates 20 messages every 10 milliseconds. Two network topologies are analysed, one representing a relatively homogeneous network layout and the other is representing a hub-and-spoke layout. Both networks are assumed to have an average diameter, the amount of time required to go between the two furthest nodes, to be 100 milliseconds.

We consider three attack scenarios. The first is when the botmaster or its sensors obtain the address, but no attack can be launched because the address has timed out of the window. This case is labelled “No Attack.” The second scenario is where the attack is successfully launched, but the attacker address slides out after 50 milliseconds. The third scenario is when the attack starts the length of the window is greater than 100 milliseconds, so no mitigation is possible. Figures 4 and 5 show the results, in terms of messages delivered, for both topologies and each of the three scenarios. The labelled data points are for the attack scenario with a window size of greater than 100 milliseconds. As can be seen, for both topologies these attacks have an immediate and dramatic impact on system goodput. For the homogeneous case the attack starts at one end of the network, while for the Hub-and-Spoke case the attack starts at some of the Hubs. For both topologies once the DRM protection mechanism kicks in network throughput is rapidly restored, even for the effective approach of Attack-Hubs-First.

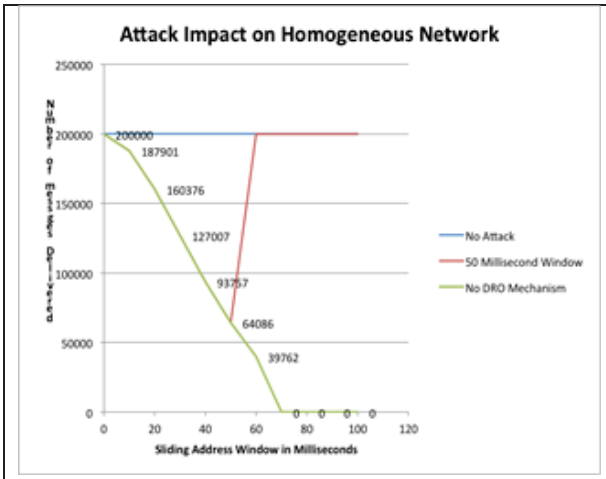


Figure 4: Results

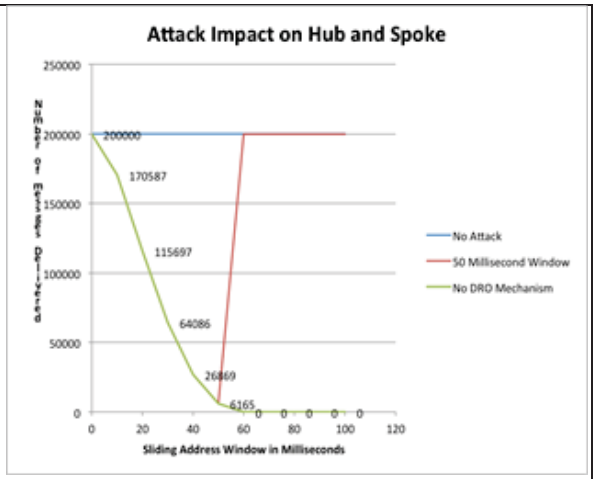


Figure 5: Results

References

Abdullah, M., and Simon, R., Analysis of Core-Assisted Routing in Opportunistic Networks, IEEE MASCOTS, Istanbul. Turkey, October 2007, pp. 387-394.

Harney, H., et al. *GSAKMP: Group secure association key management protocol*. No. RFC 4535.

Jafarian, Jafar Haadi, Ehab Al-Shaer, and Qi Duan. "Openflow random host mutation: transparent moving target defense using software defined networking." *Proceedings of the first workshop on Hot topics in software defined networks*. ACM, 2012.

Lu, Wei, and Ali A. Ghorbani. "Botnets detection based on IRC-community."*Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*.

Koponen, Teemu, et al. "A data-oriented (and beyond) network architecture."*ACM SIGCOMM Computer Communication Review*. Vol. 37. No. 4. ACM, 2007.

Ramanathan, Ram, et al. "Prioritized epidemic routing for opportunistic networks." *Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking*. ACM, 2007.

Sterbenz, J, et al, “Evaluation of network resilience, survivability, and disruption tolerance: analysis, topology generation, simulation, and experimentation,” Telecommunication Systems, February 2013, Volume 52, Issue 2, pp 705-736.

Uddin, Md Yusuf Sarwar, et al. "Making DTNs robust against spoofing attacks with localized countermeasures." *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2011.*

Increased C-Suite Recognition of Insider Threats Through Modern Technological and Strategic Mechanisms

Amie Taal[1], Jenny Le[2] and James Sherer[3]
[1]DeutscheBank AG, New York, USA
[2]UBIC NA, New York, USA
[3]BakerHostetler, New York, USA
amie33_uk@yahoo.co.uk
jle@ubicna.com
jsherer@bakerlaw.com

Abstract: A C-Suite is responsible for the actions of its organization, and this responsibility indicates that the C-Level executives may be expected to know the "in's-and-out's" of the organization to a remarkable degree of specificity. But the days when an organization's chief executive would walk the plant and point out safety or manufacturing concerns to the floor supervisor are long gone, despite the popularity of "undercover boss" type programming. It is much more likely that C-Suite executives have very little visibility into the day-to-day processes and tools their employees utilize, as the knowledge economy has increased worker specialization and employee tasks are much closer to bespoke processes. In response to this new type of working environment, employees are choosing their own tools, preferences, and applications, and integrating those into their organization's day-to-day activities. This in turn, may cause difficulties associated with top-down management of information, security governance and strategic planning. Traditional office paraphernalia—like photo frames and potted plants—are now internet of things (IOT)-enabled bring-your-own-devices (BYOD) which, quite unlike the plants pose significant challenges and risks to the organization's information technology infrastructure and create further challenges associated with insider threats, whether intentional or not. This technological revolution, its impact on worker activity and an increased possibility of intentional and unintentional insider threats is not without a silver lining. The underpinnings of the technologies that allow workers to better do their day-to-day work can also enable executives to focus on the important data points of an employee's behaviors and activities, one can then summarize and present those data points in an intelligible way to support and provide a defensible platform for corporate decision making. In this paper, we highlight some of these modern tools and explain how they may be used within the organization to look at a number of insider threats, including ongoing fraud and corporate malfeasance. We also looked at unintentional "bad practices" that, while not directed towards corporate harm may expose the organization to data breach or other leakage issues and some of the newest trends that incorporate behavioral analytics to identify those employees most likely to engage in harmful activities, before the employees have taken the first "wrong" step down that path. This paper will also examine the care with which these tools should be deployed and suggest some constraints for their operation and use. Finally, we will review how the data from these tools can best be distilled into reports appropriate for C-Suite review and utilization in support of executive decision making while still considering privacy and other employee and legal issues.

Keywords: c-suite, cybersecurity, fraud, insider threats

1. C-Suite responsibility for cybersecurity and insider threats

While executives have broad latitude under the business judgment rule when guiding an organization, this guidance may include the maintenance of cybersecurity generally. This, in turn,"ultimately involves protecting a company's confidential information and the infrastructure used to house and manage it" (Shea et al, 2015). Insider threats are among the most risky of cybersecurity threats (Miller & Maxim, 2015), and are "becoming increasingly worrisome," especially to corporate security executives (Mello, 2015b). But do C-Level executives really need to be worried about cybersecurity and related cybersecurity insider threats (CITs)? We assert that CITs should be on the proverbial radar for CEOs and those C-Level executives supporting the CEO's office. We explore this more fully below.

The C-Suite faces the investing or purchasing public, regulators and governmental agencies and key partners within the organization's ecosphere. Furthermore, public perception may be tied to CEO and executive responses to data security incidents, as seen when Target's CEO Gregg Steinhafel resigned "from all his positions" following Target's well-publicized data breach incident (Trefis Team, 2014). The equation we present is therefore straightforward: high Executive Visibility (EV) plus high levels of Executive Responsibility (ER) may lead to increased Levels of Scrutiny (LoS) and an increased weight of Public Perception (PP) (Nelson & Zeitz, 2015). Or,

***h*EV** + ***h*ER** = ↑**LoS** + ↑**PP**

Regulators in the United States, Europe, and elsewhere also focus on data security and have mentioned C-Level direction in the case of ER. For instance, on October 13, 2011, the United States' Securities and Exchange Commission (SEC) released guidance on public company disclosure obligations, directing companies to analyze cybersecurity vulnerabilities and incidents and determining associated SEC disclosure requirements if incidents represented "a material event, trend or uncertainty *that is reasonably likely to have a material effect* on the registrant's results of operations, liquidity, or financial condition or would cause reported financial information not to be necessarily indicative of future operating results or financial condition" (SEC, 2011). In 2013, the United States' Department of Homeland Security (DHS) presented five points for executives to address in managing organizational cyber risk (DHS, 2013). The import was quite clear: organizations "will face a host of cyber threats, some with severe impacts that will *require security measures that go beyond compliance*" (DHS, 2013). The advent of the 2015 EU-US Privacy Shield (replacing the Safe Harbor agreement invalidated by the *Schrems* decision of the European Court of Justice (CJEU)) also incorporates "stronger obligations on US companies to protect the personal data of EU citizens" that seek to mirror those standards already in place in the EU (Azim-Khan, 2015).

C-Level executives might address the CIT issue in order to report to the DHS, the SEC, the European Regulators, the general public, shareholders and the board. What this undertaking involves is not a clear-cut but there are a number of challenges raised when articulating exactly what might need to be done. First, CITs are not simply an information technology (IT) problem, but instead present a risk management issue that may sometimes be dealt with at the C-Suite level (Mello, 2015b). The DHS agrees, stating that such discussions regarding cyber risks (including CITs) might be elevated to the Chief Executive Officer (CEO) (DHS, 2013). Other commentators concur, holding that executive involvement in championing activities may be necessary for workable systems (Smith, 2014). Second, malicious CITs have long been considered underappreciated even by IT-specific executives (Computer Economics, 2010). If, as we assert, the C-Level may effectively address CITs, the challenge is only heightened when the C-Level must also educate the technicians upon whom the C-Level eventually rely upon to implement policy and strategic decisions.

C-Level executives may consider engaging if only for selfish reasons—they may even be primary targets of CITs. The C-Level executives in most instances hold the proverbial keys to the kingdom and some targeted attacks specifically address issues such as, "[f]rom a psychology perspective, if you were a high-profile exec[utive for an organization,] what kind of links would you click on? What are your interests?" (O'Connor & Reuille, 2016). In fact, "one of the most serious risks many corporate executives face derives from within their own organizations: disgruntled employees with real or perceived grudges" (West et al, 2016).

The final challenge we present is that, if executives are tasked with dealing with cybersecurity, they must balance that specific responsibility with other responsibilities that *already* come natively with the executive role generally—namely improving productivity, savings, and worker efficiency and satisfaction (Mehta, 2014). This is a two-faceted issue, as many C-Level executive achievements (such as meeting mission objectives and delivering business functions) may rely on information systems and the Internet. And while these CITs may share many technological characteristics, they may also represent true earth-shaking impact regardless of the organization's industry. Financial services may face account robbery or insider trading; manufacturing firms may have their intellectual property stolen; and retailers may have customer and employee data leakage (AT&T, 2015). But solving—or at least addressing—these responsibilities can work counter to the ultimate goal of a secure cyber environment where CITs are eliminated or entirely managed (McLellan et al, 2015).These responsibilities remain paramount and addressing them requires a balance we explore below.

2. C-Suite visibility into cybersecurity and insider threats

In past years, C-Level executives called up the head of IT and asked for a confirmation of existing protections. However, present-day commentators assert that C-Level executives may "no longer take for granted that their companies' assets are protected by the corporate firewall, ubiquitous virus filters and anti-spam technologies" (Boisvert, 2014). Further, the best modern technological firewalls will suffer when most insider issues themselves involve the human element, and prey on employees' natural (and helpful) tendencies to trust one another (and share information or passwords they shouldn't) (Miller & Maxim, 2015). Finally, employees are not the only vectors for insider threats—contractors and vendors with access to infrastructure or employees can also be "motivated by politics, revenge, greed" or engage in "basic corporate espionage" (AT&T, 2015).

CITs and cybersecurity generally pose a distinct and different threat for C-Level evaluation, as "two key differentiators between cybersecurity and other enterprise risks are diversity andinterdependence," (Boisvert, 2014) and organizational silo-ing tendencies. Likewise, system logs and large data streams might sap IT storage resources, the organization's IT professionals' time and be considered a drain on the organization's bottom line. However, in this specific context, they might be the only accumulated evidence of "malicious insider activity" (Parveen, 2013) or serve as the only available data to rely upon in an investigation of an incident that might take "months or years to discover" (Boisvert, 2014). Data science modeling may utilize these logs (such as Active Directory data) to evaluate and detect CITs—especially when used in conjunction with "meta user information from Human Resources or project staffing databases" to provide additional insight into a flagged user's activities (Lin 2014).This proposed change in practice will cost in the short-term, but in the data breach response context, time is also money as a failure to quickly identify an incident such as a CIT will lead to higher costs. Despite the technology challenges associated with this brave new world, not only may executive knowledge be warranted—executive support and championing activities may also be core to building mitigation programs that address CITs among other IT security issues (Smith, 2014).

3. New technologies present new cybersecurity and insider threat challenges

To set the stage for an examination of CITs, we utilize the following three-rubric classification standard: *Malicious Insiders* actively and intentionally attack; *Exploited Insiders* are misled into providing data or passwords; and *Careless Insiders* press incorrect keys, change system settings or accidentally delete critical information (Miller & Maxim, 2015). Resulting threats may presently manifest according to these typologies: *Spear Phishing* attacks, where an email spoofing fraud seeks unauthorized access to confidential data; *Watering Hole* attacks against an industry or demographic which utilize a compromised website hosting malware affecting and traveling with visitors; *Distributed Denials of Service* (*DDoS*) and *Botnets*, essentially internet traffic attacks that overwhelm or "bring down" a target site or service; *Point of Sale* (*PoS*) and *Automated Teller Machine* (*ATM*) attacks that focus on vulnerable spots within an organization's network; and others, such as *Supply Chain* attacks; *Industrial Control Systems*(Boisvert, 2014); *Encryption* and *Anonymization* tools used by employees in the workplace for professional and personal aims (Sherer et al,2015); and *Extortionware* (Boisvert, 2014).

One CIT manifestation is those insiders who "maliciously or unwittingly steal, erase, or expose sensitive data" (Miller & Maxim, 2015). These insiders—who, as noted above, may not be true "insiders"—may be motivated by a multiple factors, including revenge, money, whistleblowing, hactivisim, espionage, and business advantage (AT&T, 2015). Forms of potential CITs do not simply change according to the actors' complexities. However, while new technologies and their application create a myriad of potential CIT challenges, our space constrains us to addressingonly three exemplars:

- Modern workplace technological advancements allowing global employee interaction and work mobility. This bespoke technological application discussed further below, has made professional lives more efficient, productive, and collaborative—because the modern office therefore exists everywhere. "Everywhere-ness" also makes executive data security monitoring that much more complex and bespoke as well.
- Workforce-level changes from the influx and rapid change of technologies. The knowledge economy has increased worker specialization, making employee tasks themselves seem much closer to bespoke processes (Malone et al, 2011). Today's C-Suite executives therefore may have limited visibility into the day-to-day processes and tools their employees utilize because they simply cannot know what their employees are doing at any given point-in-time because they would have to be doing the work in order to know what, exactly, is happening.
- These bespoke work situations also involve and include the IOT and its "estimated 50 billion 'things,'" including car sensors, utility meters, and household appliances (including the afore-mentioned picture frames); cloud computing and the "explosion of mobility;" BYOD; "big data" analysis, which often utilizes "massive amounts of sensitive information" but just as often has "few built-in security safeguards;" extended supply chains, which frequently put "more data in the hands of more insiders at [other] companies;" and home health monitoring devices, where the organization is still responsible for the data collected and stored, but users remain the "hands-on" managers of that process (AT&T, 2015). This data generates "a lot of traffic," which in turn leads to a needles in a haystack approach where the "needles are playing hide-and-seek" (O'Connor & Reuille, 2016). And, as put bluntly by at least one author, "everything that is networked is hackable" (Boisvert, 2014).

Advances in technology and the availability of the same to Malicious Insiders (such as zero-day exploits) remain a long-standing problem (Schwalb, 2007). But further challenges are added by increased complexity of systems generally, including BYOD programs where each user including a vast number of potential Exploited and Careless Insiders, chooses his or her own technology of choice. The organization is left attempting to deal with multiple technologies, risk profiles and a myriad of end-node considerations (McLellan et al, 2015). This also introduces the beginning of other concerns, such as Shadow or "credit card" IT, where employees utilize hardware and software as part of their jobs without any notification to corporate IT at all (Walters, 2013) and such utilization may contribute to around 40% of all IT misuse (Boisvert, 2014). Increased proliferation of IOT may also cause issues for traditional means of shoring up security, such as applying software patches to combat known security issues (Mello, 2015b).

Increased numbers of IT assets is only part of the story, the actual numbers only add to and are compounded by the stress on the IT infrastructure from the users' use of the IT assets (including the addition of those Shadow IT assets). Engineers tasked to deal with CIT challenges must therefore factor technological and behavioral elements associated with the IT environment as a whole; these include such specifics as "[b]ackground jobs, RAM, CPUs, Parasites, [and] Hardware Failures" (O'Connor & Reuille, 2016).

4. New cybersecurity and insider threat challenges beget new technologies – and new solutions

New technologies signify hope for meeting CITs, but do not present a panacea that will eliminate them. Time and time again, writers reiterate that "cyber threats are as much about people as technology" (Boisvert, 2014); this is doubly true for CITs. Technologies that incorporate new types of analysis using expanded data pictures (such as the big data issues discussed above) are key to incorporating this wisdom into an executive strategy addressing the CIT threat. Experience with these types of CIT issues has led to a more certain picture of the behavioral indicators associated with CIT activities, including red flags many workaholics might raise inadvertently (including "remotely access[ing] the computer network while onvacation, sick leave, or at other odd times" (DOJ, 2011)). Using behavioral or "stream" analytics to mine and consider big data has therefore appropriately been part of organizations' considerations regarding CIT for a number of years now (Parveen, 2013).

While trust, as discussed earlier poses a significant threat to organizations within the context of insider threats, it cannot—and should not—be eliminated. Instead, executives must "find the right balance between employee enablement and control, while holding employees accountable for their actions" (Miller & Maxim, 2015). But if CIT mitigation is the true end-goal, these types of recommendations should be rooted in a *Moral View*, which aims to achieve "good ends from the perspective of the big picture and the long view" rather than the *Moralistic View*, which generally takes "a short-term view, never mind the bigger picture" (Charney, 2014). This may include the use of new understandings as research continues to address the identities and motivations of Malicious Insiders, in ways that can be translated into big data tool analytics (Charney, 2010; Charney, 2014). Tools that can utilize ensemble based learning algorithms that maintain dictionaries of "repetitive sequences found throughout dynamic data streams of unbounded length to identify anomalies" leading, to models for common behavior sequences and increased classification accuracy for "data streams containing insider threat anomalies" (Parveen, 2013).

Approaches to CIT mitigation arise in a variety of different situations. In addition to the big data behavioral modeling and response discussed above, a novel method to detect Secure Sockets Layer (SSL) man-in-the-middle attacks was presented by Huang et al (2013) where in order to determine whether a SSL connection was being intercepted, the researchers observed the server's certificate from the client's perspective. Intuitively, if the client actually received a server certificate that did not exactly match the website's legitimate certificate, there would be direct evidence that the client's connection must have been tampered with. Through this detection method, the researchers collected 845 forged certificates from real-world clients connecting to Facebook's SSL servers and were able to survey the characteristics of the forged certificate chains, including the certificate chain sizes, certificate chain depths, public key sizes and examined the subject names and the issuer names of the forged certificates (Huang et al, 2013). The researchers were then able to categorize the certificate issuers of forged certificates into antivirus, firewalls, parental control software, adware and malware. While many did not represent suspicious activity, the researchers noted some instances of forged certificates of questionable origin that would require further investigation. The data suggest that similar mechanisms can help browsers to possibly

detect forged certificates based on size characteristics, such as checking whether the certificate chain depth is larger than one; they may also contribute to the greater good of white-hat security efforts by providing information for blacklist sites (O'Connor & Reuille, 2016).

While anomalies may signify some kind of danger, context is helpful in evaluating the actual threat involved. Other big data analytic solutions attempt to "create a context around everything connected to the network – users, accounts and devices" in order to see if a user is engaging in activity that might seem "abnormal" within the context of the entire organization, but if normal within a defined peer group, that concern may evaporate (Mello, 2015a). These new types of computer- and big data-applications rely less on traditional forms of supervised learning that require "well-balanced training data," (Parveen, 2013) and instead focus on unsupervised learning algorithms already churning through many different data sets for different applications, including sensitive legal document response requirements (Zeinoun et al, 2011).

Executives may consider that other insider threats can include ongoing fraud and corporate malfeasance—unintentional "bad practices" that, while not directed towards corporate harm, may expose the organization to data breach or other leakage issues. For example, some software vendors have tools implemented in various industries, where artificial intelligence is used to drive an email auditing process through automated learning that is able to identify "at risk" email communications (UBIC, 2016). This technology is able to identify risky behavior and communication content to address potential cartel and antitrust activity, intellectual property theft and general corporate malfeasance. In addition to being able to detect current activity, the technology is able to identify three stages of activity: development, preparation and execution. Taken together, these phases allow companies to both predict the risk of certain behavior as well as mobilize early to mitigate the activity before it becomes a reality (UBIC, 2016); other solutions offer similar "volatile data (e.g. currently running processes, loaded DLLs, memory)" solutions for risk and employee monitoring (AccessData, 2008).

The analysis of hundreds of insider threat case studies has led to additional, less absolute options for dealing with insider threats (Silowash et al, 2012). Charney followed his original insider spy psychology studies with what he termed an approach to allow insider threats (primarily government or military, but the same approach would apply) a "safe exit pathway" for those insiders "prepared to embark upon [a] reconciliation track, and accepting its demanding conditional provisions" (Charney, 2014). Other commentators note that this may include the organization's employee assistance program (EAP) if one exists (Silowash et al, 2012). Charney's proposal both considers the human element of insider attacks, as Miller & Maxim (2014) suggest, but also provides a measured rather than absolute approach to dealing with CITs. Silowash (2012) also incorporates the human element into their proposed nineteen best practices for the prevention and detection of insider threats, especially when directing organizations to "[e]stablish a baseline of normal network device behavior."

Many of these approaches focus on the lack of definition or perceptions regarding CIT-related activities. The DOJ highlights these challenge when identifying related organizational factors, including "[u]ndefined policies regarding working from home on projects of a sensitive or proprietary nature" as well as perceptions "that security is lax and the consequences for theft are minimal or non-existent" (DOJ, 2011). These considerations of new behavioral research conclusions are helpful for the application of new monitoring applications. Additional technologies now focus on such issues as Autonomous System Number (ASN) whitelisting, where domains exhibiting fraudulent behavior are sometimes observed as being hosted on ASNs unassociated with the organization they're spoofing (O'Connor & Reuille, 2016).

Natural Language Processing (NLP) tools are not necessarily "new," but new approaches using tried-and-true technology can contribute to more effective approaches to identifying CITs or even curbing opportunities for related employee action (O'Connor & Reuille, 2016). These also include a number of tools used traditionally by eDiscovery or organizational Internal Investigation teams that take advantage of the common language of data and systems, and consider a number of varied techniques from different disciplines when investigating the story of an incident or issue (Taal, 2015). Finally, a mixture of psychology and technical solutions should lead to a combination of detection models that incorporate the right blend of technologies that incorporate both new and tried-and-tested big data analytic techniques, as well as a firm appreciation for the human psychology and trust that underlie the entire organizational dynamic.

5. Conclusion

To support executive decision making, as part of a tactical first-step approach, C-Level executives tasked with or considering responses to potential CITs might begin with how (and whether) the organization maps to benchmarking standards. Asking IT or others whether the organization is compliant with NIST: RA-1, RA-3, PL-1,PL-4, PM-9, and PS-8 (to name just a few); CERT-RMM; or the security policies and standards for ISO 27002 – 8.2.2, 10.10.2, 10.10.4,and 15.2.1 (again, as a sampling) might start a dialogue and could spur immediate action on the part of those responsible parties if warranted (Silowash et al,2012).

This approach could incorporate "a broad approach to allow an organization to carefully manage its identities, access and data, from identity management, to governance, privileged identity management and data protection"(Miller & Maxim, 2015). This might include such tenets as strengthening the organization's security foundation; making security everyone's responsibility and thereby at least informing Exploited and Careless Insiders that they may be, even if unconsciously, part of the problem; breaking down organizational silos where hackers might otherwise "count on bureaucratic inefficiency and barriers between groups;" and invest in behavioral analytics, utilizing some of the big data tools discussed above (while still recognizing that these tools become additional repositories of sensitive data that require a separate security analysis) (AT&T, 2015). Using big data analytics at this point may involve "purpose-built data mining, correlation, enrichment, and analytics" in order to detect "not only users with high risk identity profiles but also high-risk activity, access, and events in [an] organization" associated with CIT (Securonix, 2015).

Finally, as Charney (2014) notes, spy (and by extension, CIT) prevention entails "measures designed to help a person climb back [into a productive professional capacity] when they're in danger of getting overwhelmed by the stresses that are battering them." Training and related considerations that address these concerns could be addressed early within employees' (and others') professional relationship with the organization. The identified disciplines of "employment, privacy, employee benefits and executive compensation are indispensable to building an effective program to reduce the risk of cybersecurity incidents occurring, in the first instance and to respond effectively once a breach is suspected or has occurred, all without running afoul of applicable laws" (Shea et al, 2015). A C-Suite executive might therefore address the less attractive side of the human element when planning strategy, but include forgiveness in the approach. While trust was identified earlier as a cause for concern and a weakness within the process (Miller & Maxim, 2015), a more considered approach might follow a "trust but verify" model (Shea et al, 2015). This is consistent with the "risk-managed approach to information access" championed by a number of commentators (Boisvert, 2014), where the C-Level must trust their and their employees' understanding of the CIT, and that the group is working in harmony to enable the organization as a whole to constructively address CIT issues generally.

Disclaimer: The views expressed herein are solely those of the authors, should not be attributed to their places of employment, colleagues, or clients, and do not constitute solicitation or the provision of legal or security advice.

References

AccessData (2008) AccessData Enterprise, System Security Overview, Whitepaper.

AT&T (2015) Insider Threats, AT&T Cybersecurity Insights.

Azim-Khan, R. (2016) The EU and US data transfer agreement: perhaps a shield, but no silver bullet, TechCity News (February 7, 2016).

Boisvert, R. (2014) What every CEO needs to know about cybersecurity: A background paper, I-SEC Integrated Strategies Whitepaper.

Charney, D. (2010) "True Psychology of the Insider Spy", Intelligencer: Journal of U.S. Intelligence Studies, Fall/Winter 2010, 47-54.

Charney, D. (2014), NOIR: A White Paper, NOIR4USA.ORG Whitepaper.

Computer Economics (2010) Malicious Insider Threats Greater than Most IT Executives Think, Malicious Insider Threats report excerpt.

Department of Justice (DOJ) Federal Bureau of Investigation (FBI) (2011) The Insider Threat – An introduction to detecting and deterring an insider spy, FBI Brochure.

Huang, L.-S., Rice, A., Ellingsen, E., and Jackson, C. (2014) Analyzing forged ssl certificates in the wild, IEEE Symposium on Security and Privacy.

Lin, D. (2014) A Data Science Approach to Detecting Insider Security Threats, Pivotal Data Science Blog (June 12, 2014).

Malone, T., Laubacher, R., and Johns, T. (2011) "The Big Idea: The Age of Hyperspecialization", Harvard Business Review, July-August 2011 Issue.

McLellan, M., Sherer, J., and Fedeles, E. (2015) "Wherever You Go, There You Are (With Your Mobile Device): An Examination of Privacy Risks and Legal Complexities Associated with Cross-Border 'Bring Your Own Device'Programs", Richmond Journal of Law and Technology, Vol. 21 No. 11.

Mehta, A. (2014) "'Bring Your Own Glass:' The Privacy Implications of Google Glass in the Workplace", John Marshall Journal of Information Tech and Privacy Law, Vol. 30.

Mello, J. (2015a) Big Data Analytics Fights Insider Threats, TechNews World (May 13, 2015).

Mello, J. (2015b) Security Execs Sweat Insider Threats, TechNews World (December 31, 2015).

Miller, R. and Maxim, M. (2015) I Have to Trust Someone….Don't I? CA Technologies White Paper.

Nelson, L.B. and Zeitz, J. (2014) How Law and PR Can Work Together In High-Profile Cases, Law360 (April 10, 2014).

O'Connor, J. and Reuille, T. (2016) The Security Wolf of Wall Street: Fighting Crime with High-Frequency Classification and Natural Language Processing, OpenDNS Presentation (January 2016).

Parveen, P. (2013) Evolving Insider Threat Detection Using Stream Analytics and Big Data, Dissertation for Doctor of Philosophy in Computer Science, The University of Texas at Dallas (December 2013).

Schwalb, M. (2007) "Exploit Derivatives & National Security", Yale Journal of Law and Technology, Vol. 9, No. 1, 162-192.

Securonix (2015) Bringing Clarity to Insider Threat, Securonix Insider Threat Management Blog.

Shea, R., Burke, L., Fein, A., and Bracebridge, C. (2015) Insider threats to cybersecurity – An HR legal perspective, Inside Counsel (March 12, 2015).

Sherer, J., Le, J., and Taal, A. (2015)Big Data Discovery, Privacy, and the Application of Differential Privacy Mechanisms, ICAIL 2015 Workshop on Using Machine Learning and Other Advanced Techniques to Address Legal Problems in E-Discovery and Information Governance ("DESI VI Workshop"); The Computer & Internet Lawyer, Vol. 32 No. 7.

Silowash, G., Cappelli, D., Moore, A., Trzeciak, R., Shimeall, T., and Flynn, L. (2012) "Common Sense Guide to Mitigating Insider Threats", Carnegie Mellon Software Engineering Institute Technical Report CMU/SEI-2012-TR-012, 4th Ed.

Smith, J. (2015) Mitigating malicious insider cyber threat, Technical Report RHUL-MA-2015-12, Information Security Group, Royal Holloway University of London.

Taal, A., Le, J., and Sherer, J. (2015) A Consideration of eDiscovery Technologies for Internal Investigations, Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security, 10th International Conference, ICGS3.

Trefis Team (2014) Target's CEO Steps Down Following The Massive Data Breach And Canadian Debacle, Forbes (May 8, 2014).

UBIC (2015) Lit I View Email Auditor, AI-Based Data Analysis Platform Product Sheet.

United States Department of Homeland Security (DHS) (2013) Cybersecurity Questions for CEOs, US-CERT Security Publications.

United States Securities and Exchange Commission (SEC) Division of Corporation Finance (2011) Cybersecurity, CF Disclosure Guidance: Topic No. 2 (October 13, 2011).

Walters, R. (2013)Bringing IT out of the Shadows, Network Security, Vol. 2013, No. 4, 5-11.

West, C., O'Hara, R., and Jantzen, B. (2016) Corporate Executive Protection: 7 Trends to Watch in 2016, AS Solution Blog.

Zeinoun, P., Laliberte, A., Puzicha, J., Sklar, H., and Carpenter, C. (2013) Recommind at TREC 2011 Legal Track, Submitted Paper, Text Retrieval Conference (TREC).

Work in Progress Paper

Creation of Specific Flow-Based Training Data Sets for Usage Behaviour Classification

Florian Otto[1], Markus Ring[1], Dieter Landes[1] and Andreas Hotho[2]
[1]Coburg University of Applied Sciences, Coburg, Germany
[2]University of Würzburg, Germany
florian.otto@hs-coburg.de
markus.ring@hs-coburg.de
dieter.landes@hs-coburg.de
hotho@informatik.uni-wuerzburg.de

Abstract: The majority of security methods for computer networks rely on well-known signatures for detecting malware and attacks. Consequently these methods often fail to detect novel or obfuscated attacks in monitoring data. Usage behaviour classification and anomaly detection seem to be promising approaches to address these problems. In this context, the development of sophisticated classifiers allowing to sift monitored data caused by harmless behaviour patterns can greatly improve the results of downstream anomaly detection methods. The former rely on a valid ground truth, defining harmless usage behaviour sufficiently. Recognising behaviour patterns also may facilitate the detection of insider threats, if single employees behave not as expected. We propose a workflow enabling us to create labeled flow-based data sets containing information about real usage behaviour. Within this workflow, real humans use virtual machines to work on specific tasks and simultaneously log their activities. In addition, the virtual machines log processes inducing network connections. Both logs are then used to attach labels to the monitoring data, to enrich it with information about the corresponding usage behaviour. We describe the process with an example scenario of an adapted computer pool of our university. Finally, the resulting data set is briefly discussed.

Keywords: behaviour classification, data set generation, intrusion detection, anomaly detection, flow-based data

1. Introduction

Network Intrusion Detection Systems (NIDS) are used to ensure network security and to encounter attacks. NIDS can be distinguished in misuse detection and anomaly detection (Sommer and Vern, 2010). Misuse detection searches for well-known signatures of malware or attacks in network monitoring data (Giacinto et al., 2008). Signatures have to be constantly updated, since malware and attacks are constantly developed, refined and obfuscated (Giacinto et al., 2008). In contrast, anomaly detection tries to distinguish normal from abnormal behaviour. Sufficient recognition of harmless behaviour facilitates detection of malware, attacks and misuse, considering those lead to mutations within the data. Corresponding methods e.g. classifiers rely on representative training data sets containing a valid ground truth. Due to privacy concerns and difficulties to label network monitoring data, few publicly available data sets exist (Sommer and Vern, 2010).

To address this problem, we propose a workflow which allows to easily built networks and gather monitoring data within pre-defined scenarios. The networks are built in virtual environments that allow real humans to work with and to monitor the network in a controlled environment. Since we focus the enrichment of the data with information about corresponding usage behaviour, the users log their activities manually. Additionally, the virtual hosts log processes inducing network traffic. Our environment allows to adapt networks, which can than be used for specific scenarios without interfering with productive systems. Pre-defined scenarios may be defined as the normal usage of the real counterpart network to gather harmless behaviour, or special situations which may appear more seldom. Penetrations tests or attacks can also be performed or malware could be installed to see effects within the monitored data.

Our main contribution is a way to create network traffic of adapted productive networks in a controlled environment and to enrich it with information about real usage behaviour.

2. Related work

Intrusion detection data sets can be distinguished in packet-based, flow-based and application-based. Since the proposed data set is flow-based, the following review considers only public available flow-based data sets.

One of the first labeled flow-based data sets for intrusion detection was contributed by Sperotto et al. (2009). Flows were collected from a real honeypot with HTTP, SSH and FTP services. The log files of the services were used to label the corresponding flows. The resulting data set contains over 14 million flows, but most of them are suspicious due to missing further background traffic.

Shiravi et al. (2012) describe a systematic approach to generate adapted data sets for intrusion detection. Their basic idea is to describe profiles which contain detailed descriptions of normal user activities and attacks. Based on these profiles, they generated and published the ISCX data set.

Another publicly available flow-based data set is CTU 13 Malware (García et al., 2014). In this data set, 13 different botnet scenarios are mixed with background traffic. The data is labeled based on the IP-addresses used by the botnets.

In contrast to the above, we enrich our data set with information about real user behaviour based on manually provided activity protocols and with information about processes inducing network traffic.

3. Data set generation

3.1 Toolset

Our approach is based on OpenStack which allows the creation and management of virtual networks and virtual machines. Within this environment we are able to adapt existing networks or subnetworks of organisations and to record their corresponding network traffic. Real users which work on the virtual hosts are connected via remote connections that are routed over a single special host, enabling us to easily sift traffic caused by the remote connections.

3.2 Flow-based monitoring

Widespread standards of flow-based monitoring data are Netflow and Internet Protocol Flow Information Export (IPFIX). Flows represent connections between two network components and are mostly identified by the default five-tuple (Protocol, Source- and Destination Address, Source- and Destination Port) (Kim et al., 2004). Flows encapsulate different metrics, e.g. number of sent packets and bytes, the duration and used TCP-Flags. The encapsulation of flow-based monitoring significantly reduces the amount of data to be stored in comparison with full-packet-monitoring. Additionally, no contents of communication are stored leading to less privacy concerns.

3.3 Labelling

Our goal is to create monitoring data sets enriched with information about usage behaviour. Therefore, we use a two-fold labelling strategy: Firstly, we monitor processes and applications using sockets on the host machines. This allows us to label the flows with applications which induced them. Secondly, human users are instructed to log their activities. Informal and detailed logs would result in complicated extraction information. Although concise terms contain less information they are easier to evaluate. Further, participants should not be distracted too much by writing protocols, since this would distort their behaviour. In due we provide a small extensible taxonomy of terms with short descriptions, which the participants could select. This labelling procedure is quick and flexible enough to adapt specific behaviours. Activities may overlap since it is likely to do things in parallel with computers e.g. searching for information while communicating. We attempt to create realistic data sets, so we don't restrict the usage to one activity at a time. However, this may lead to multiple activity labels per flow.

4. Scenario

As an example scenario, we adapted student workplaces of a computer pool at the university. We prepared six virtual machines with Ubuntu 14.04 LTS. Within a two hour timeframe, the students were asked to do research, work on current projects or work on online tutorials about programming languages. The experimentees were also asked to perform tasks they do in leisure time e.g. streaming videos, using social networks, etc.

5. Resulting data set

In the experiment five of the six hosts were used by human participants. One of the hosts was active but not used (*User B* in Figure 1). An overview of the activities during the experiment is given in Figure 1.

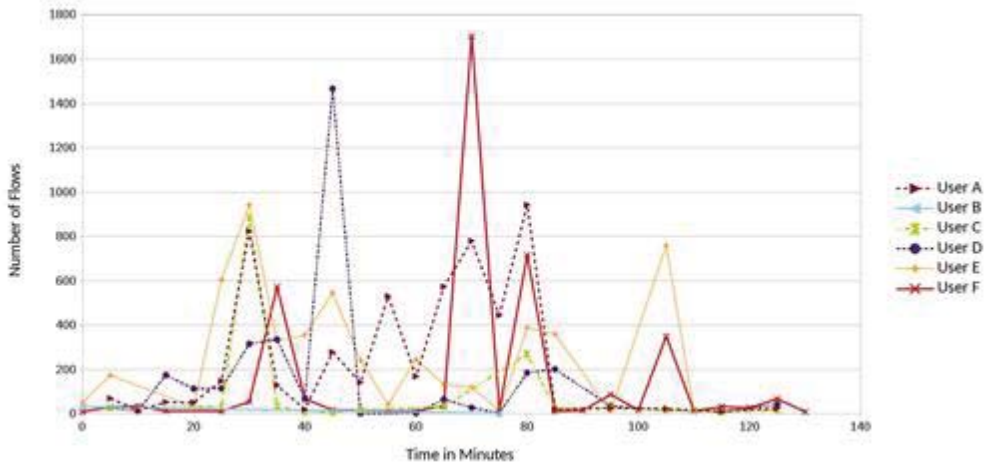


Figure 1: User activity represented as number of flows within 5 minute intervals

The resulting data set contains *19,222* bidirectional flows. The majority, *19,180* (*99.78 %*) of the flows were induced by the hosts used by the participants.

Table 1: Process labels within the dataset

Process Name	Number of Flows
firefox	7,001
plugin-container	98
thunderbird	7
ssh	5
unknown	12,111

As shown in Table 1, *7,111* (*36.99*%) flows are labeled unambiguous with host processes. For the larger proportion of flows no process could be identified (*unknown)*. However, *11,306* (*93.35*%) of these can be lead back to the local domain name service of the environment. Apart from that almost all user activity which can be seen in the flow data is induced by the web-browser, since Firefox and its plugin-container dominate the whole data set.

Table 2: The top ten used terms describing activities during the experiment

Activity	Number of flows	Percentage
surf the web	7,750	40.32
research	6,015	31.29
IRC (webclient)	3,731	19.41
online gaming	2,447	12.73
listening music	2,038	10.60
reading news	1,805	9.39
file transfer	1,199	6.24
online tutorials	884	4.60
videostreaming	538	2.80
mailclient	274	1.43

Table 2 shows the top ten different labels for activities which were chosen by the participants during the experiment. Note that the labels overlap and exceed the total number of flows. The two most frequent labels are *surf the web*, which was described as using the web for private interests and *research* meaning the aimed search for information about specific topics. One of the participants used the internet relay chat (IRC) for communication within a browser application. The participant used that in the usual way having the application active for almost the whole time, even if it was only rarely used. Due to that, almost every flow of that participant carries that label, which explains the high proportion.

The resulting data set shows that the label for processes introduces little information, since almost all user activity results in flows induced by the web-browser. Since many network services next to the classical presentation of internet pages, e.g. video-streaming, music-streaming or communication can be used within browsers, we expect similar outcomes in other scenarios. Nevertheless, if sophisticated networks are adapted where special software is used or if participants can use their own preferred software, this label will carry more information.

6. Future work

An open question is if the usage of virtual environments lead to significant differences in the monitoring data compared to data which is created by physical components. The usage of specific applications or tasks should not vary, since there are few differences for human users. However variations in temporal behaviour of the components or other effects are not ruled out.

Further, we plan to adapt more modular networks to create a wide range of data sets describing different scenarios. This includes performing penetration tests, infecting hosts with malware and simulating insider threats.

Finally, we want to use these data sets to train sophisticated classifiers for distinguishing between normal and abnormal usage behaviour in productive networks.

7. Conclusion

We proposed a workflow and toolset enabling us to create labeled flow-based training data sets for usage behaviour classification. Therefore an OpenStack environment is used to adapt networks and to perform scenarios in which networks can be analysed without interfering with their real counterparts.

We focus the enrichment of the data with information about corresponding usage behaviour, which is why real users participating in scenarios log their activities based on a simple taxonomy.

We briefly discussed an example data set which we generated in a small scenario.

References

García, S., Grill, M., Stiborek, J., & Zunino, A. (2014) "An Empirical Comparison of Botnet Detection Methods", *Computers & Security*, Vol. 45, pp 100-123.

Giacinto, G., Perdisci, R., Del Rio, M., and Roli, F. (2008) "Intrusion Detection in Computer Networks by a Modular Ensemble of One-Class Classifiers", *Information Fusion*, Vol. 9, No. 1, pp 69-82.

Kim, A. S., Kong, H. J., Hong, S. C., Chung, S. H., and Hong, J. W. (2004) "A Flow-based Method for Abnormal Network Traffic Detection", *Network operations and management symposium (NOMS)*, IEEE/IFIP, Vol. 1, pp 599-612.

Shiravi, A., Shiravi, H., Tavallaee, M., and Ghorbani, A. A. (2012) "Toward developing a systematic approach to generate benchmark datasets for intrusion detection", *Computers & Security*, Vol. 31, No. 3, pp 357-374.

Sommer, R. and Vern, P. (2010) "Outside the Closed World: On Using Machine Learning For Network Intrusion Detection", *Security and Privacy (SP)*, 2010 IEEE Symposium on, IEEE, pp 305-316.

Sperotto, A., Sadre, R., Van Vliet, F., and Pras, A. (2009) "A Labeled Data Set For Flow-based Intrusion Detection", *Proc. of the 9th IEEE Int. Workshop on IP Operations and Management (IPOM)*, Springer, pp 39-50.

Patterns of Bureaucratic Politics Related to Commercial Military Service Providers

Mikko Räkköläinen
University of Tampere, Finland
Rakkolainen.Mikko.J@student.uta

Abstract: This article explores the bureaucratic politics by which commercial military service providers (CMSPs) influence the government agencies employing them. The outsourcing of functions previously seen as the domain of government agencies has expanded rapidly in the last 25 years. Many armed forces have become reliant on commercial support. This reliance and the integration of CMSPs into the planning and execution gives them an opportunity to influence operations. In this article, administrative science theory is applied to past studies on the phenomenon to uncover results relevant to security studies. The work is based on the bureaucratic policy model of foreign policy formation. It examines foreign and security policy as the result of power games within the state apparatus. Ministries, agencies and service branches all have priorities, procedures and goals which they promote and protect. The resulting compromises determine policy. While the security policy apparatus in the 21st century has expanded to include numerous non-state actors, CMSPs among them, the approach remains valid. Analyzing various incidents in which CMSPs have influenced their employers' operations through the framework of organizational power proposed by Mintzberg makes it possible to conceptualize the basis of their power, the means by which it is employed and the goals which they aim to achieve. The results of this article allow for a better understanding of CMSPs as actors within the state foreign and security policy machinery, highlighting their differences in comparison to traditional armed forces. Furthermore, they help to understand the impact the increased participation of the commercial sector has on military and crisis management operations. The article also provides a means of evaluating assumptions presented in previous research of CMSPs and guides further research into their political consequences.

Keywords: decision making, bureaucratic politics, commercial military service providers, private military companies

1. Introduction

In this article I explore the ways in which commercial military service providers (CMSPs) impact the operational decision making of military and civil crisis management operations. I approach the phenomenon through the lens of the bureaucratic politics model of foreign policy and borrow theoretical concepts from administrative science. The application of these tools to a sample of cases found in previous research concerning CMSPs allows for a better understanding of the practical impact of military outsourcing and provides proof of concept for further research.

Since the end of the Cold war, states have outsourced functions related to military operations and foreign policy execution (Heinecken 2014, pp.629–630). Today state agencies purchase services as varied as armed protection, cyber security and training. As government agencies have sought to minimize costs by downsizing, they have become dependent on CMSPs for functions critical to their work (Ibid., p.632).

Academic research into the commercial military service industry (CMSI) has also grown. Beyond describing and categorizing CMSPs (e.g. Singer 2003) research was initially guided by themes highlighted by public discourse, such as civil-military relations, the state monopoly on violence and regulation (e.g. Avant 2005). Recently the field has expanded to cover questions such as the financial and political impact of military outsourcing (e.g. Krahmann 2010) and to encompass more empirical research (e.g. Kelty & Bierman 2013).

The study of CMSI has thus far somewhat ignored the operational level on which the companies carry out their work, largely due to difficulties in gathering data. While previous research contains numerous instances in which the impact of CMSPs on operational decision making has been noted, a systematic examination is still lacking. This study aims to in part fill this gap by examining data from previous research through a framework geared towards uncovering forms of organizational power.

Next, I will overview the theoretical framework. After that the material used in the research will be outlined. Finally, I will present and discuss results and future research avenues.

2. Theoretical framework

To identify key actors, concepts and relationships, a theoretical framework is required. This article employes the bureaucratic politics model of policy formation for this purpose. This model is suited for the research question, as it is concentrated on the policy impact of decision making processes. Additionally typologies borrowed from administrative science are also employed. These allow for a systematic analysis of individual incidents of organizational power use, which can then be understood in relation to their wider political ramifications.

The bureaucratic politics model focuses on the agencies and agents involved in the policy process. These range from ministries to service branches and from presidential advisors to unit commanders on the ground. Foreign policy is understood as a resultant of the combined actions of these different agents. These actions are motivated by differing bureaucratic roles. Agents prioritize the issues they're responsible for, promote the solutions they provide and follow their own standard operating procedures. Agents also have an interest in maximizing their material and political resources. (Allison 1971, pp.162–181; Hudson 2012, pp.20–21)

The bureaucratic politics model has been applied to security studies in the past, notably in Allison's classic analysis of the Cuban missile crisis (1971; also Puukka 2005). In previous research only government agencies have been viewed as agents suitable for study. While CMSPs differ from government agencies in many ways, there is no *a priori* reason to exclude them. As Allison (ibid., p.164) notes, "[i]ndividuals become players in the national security game by occupying a position that is hooked on to the major channels for producing action on national security issues". While many of these positions have been outsourced, they arguably still need to be included in the model to maintain its explanatory power.

A more detailed analytical tool is also needed to understand the actions players take. Such a tool is presented by Mintzberg in *Power in and Around Organizations* (1983). Developed primarily for analysis of commercial entities, the work is equally applicable to the public-private-network forming the modern foreign policy apparatus. Defining power as "the capacity to effect organizational outcomes" (Ibid., p.4), Mintzberg goes on to lay out several typologies related to it. Organizational power may be based on a number of different factors. Within the organization, four systems of power operate and vie for primacy. As different players employ these systems and come into conflict in political games, structural factors lead these games to take certain typical forms.

Mintzberg's work makes it possible to examine cases in which a CMSPs have employed organizational power and determine on what the power was based, how it was employed and what was the type of resulting power game. While taken individually, these observations represent administrative behavior. However, the resulting patterns typical to CMSPs make it possible to connect the observations to broader issues related to foreign policy execution.

The bureaucratic politics model thus informs the concepts and relationships the present study concentrates on, while the Mintzberg's theories allow data to be categorized and structured. Next I will turn to the research material.

3. Research material

The material examined in the present article consists of previous academic works involving CMSPs. From this sample, individual instances of CMSPs employing the organizational power power they have as members of the foreign and security policy apparatus are identified. These instances are then categorized according to Mintzberg's typologies, discussed above.

This type of secondary analysis has limitations. The sample is likely to overreport readily detectable or anomalous incidents (Price & Ball 2014). They may also not report sufficient detail and context to allow for full categorization within the theoretical framework. Despite these factors, indications of the systemic tendencies originating from the common attributes of CMSPs can be uncovered from the data to an useful extent. (Yorke 2011)

The sample contains works published from 2010 to 2015. The peer reviewed articles were identified from the reference databases Web of Science and SCOPUS, by applying the search terms "private military company" and "private security company", both being equally employed in literature. From the results book reviews, articles

not written in English and works related to domestic security were excluded. By applying this criteria, a sample of 34 articles was identified. In addition, the sample includes 10 books. These works have been selected to cover the phenomenon from numerous viewpoints within the field of security studies to maximize the number of cases to be found.

The present article thus makes use of an extensive sample of previous studies, while acknowledging the biases secondary analysis introduces. As the analysis of this sample has not yet begun in earnest, no preliminary results can be reported at this time. However, probable results as well as their application in further research will be discussed next.

4. Conclusions

To conclude, the present article seeks to systematically analyze incidents in which CMSPs have employed organizational power to impact the military and crisis management operations they have supported. These incidents have been collected from a sample of previous studies on the CMSI and analyzed using a series of typologies proposed by Mintzberg. The result is an understanding of basis of power, as well as the ends and means of using it, typical to CMSPs. At the time of writing the analysis of the material has not begun. However, previous research suggests certain tendencies.

An overview of the assumptions presented in previous CMSI research indicates that on the operational level the power of CMSPs is primarily based on the fact that state agencies are often dependent on them for mission critical resources and skills (de Nevers 2016, pp.171–173). This should lead them to maximize the influence of their expert knowledge (Mintzberg 1983, pp.163–170, 198–200) CMSPs have also been noted to gain substantial formal authority in relation to weak states as they carry out security sector reform programs (McFate 2016). How these tendencies translates into operational decision making behavior is, however, currently unknown.

As secondary analysis, the present work is tied to the limitations of the research material. It does, however, provide indications about forms of organizational power use typical to CMSPs. This understanding is important to developing the cooperation between state and commercial actors in crisis management operations, and to the larger discourse regarding military outsourcing. The results can also inform primary source CMSI research employing the bureaucratic politics framework regarding the most promising avenues of inquiry. As it is my intent to carry out such case studies during my PhD work, this article will be the first stepping stone of my research project.

References

Allison, G.T., 1971. *Essence of Decision: Explaining the Cuban Missile Crisis*, Boston: Little, Brown and Company.

Avant, D., 2005. *The Market of Force: The Consequences of Privatizing Security*, Cambridge: Cambridge University Press.

Heinecken, L., 2014. Outsourcing Public Security: The unforseen consequences for the Military Profession. *Armed Forces & Society*, 40(4), pp.625–646.

Hudson, V.M., 2012. The History and Evolution fo Foreign Policy Analysis. In S. Smith, A. Hadfield, & T. Dunne, eds. *Foreign Policy: Theories, Actors, Cases*. Oxford: Oxford University Press, pp. 13–34.

Kelty, R. & Bierman, A., 2013. Ambivalence on the Front Lines: Perceptions of Contractors in Iraq and Afghanistan. *Armed Forces & Society*, 39(1), pp.5–27.

Krahmann, E., 2010. *States, Citizens and the Privatization of Security*, Cambridge: Cambridge University Press.

McFate, S., 2016. PMSCs in International Security Sector Reform. In R. Abrahamsen & A. Leander, eds. *Routledge Handbook of Private Security Studies*. New York: Routledge Handbooks, pp. 118–127.

Mintzberg, H., 1983. *Power In and Around Organizations*, Engelwood Cliffs: Prentice-Hall, Inc.

de Nevers, R., 2016. Private Security's Role in Shaping US Foreign Policy. In R. Abrahamsen & A. Leander, eds. *Routledge Handbook of Private Security Studies*. London: Routledge Handbooks, pp. 168–176.

Price, M. & Ball, P., 2014. Selection bias and the statistical patterns of mortality in conflict. *The SAIS Review of International Affairs*, 31(1), pp.9–20.

Puukka, I., 2005. *Valtapelit hallinnossa: Tapaustutkimus sotilaskulttuurin puolustuksesta puolustushallinnon uudistamisessa*, Tampere: Kustannus Oy Suomen Mies.

Singer, P.W., 2003. *Corporate Warriors: The Rise of the Privatized Military Industry*, London: Cornell University Press.

Yorke, M., 2011. Analysing existing datasets: some considerations arising from practical experience. *International Journal of Research & Method in Education*, 34(3), pp.255–267.

ECCWS 2017

Dublin
Ireland

16th European Conference on Cyber Warfare & Security

University College Dublin
Dublin, Ireland
July 2017

For further information contact
info@academic-conferences.org
or telephone
+44-(0)-118-972-4148