Clustering Geo-Indistinguishability for Privacy of Continuous Location Traces

Mariana Cunha
CISUC, Dep. of Informatics Engineering
University of Coimbra
Coimbra, Portugal
mccunha@dei.uc.pt

Ricardo Mendes
CISUC, Dep. of Informatics Engineering
University of Coimbra
Coimbra, Portugal
rscmendes@dei.uc.pt

João P. Vilela
CISUC, Dep. of Informatics Engineering
University of Coimbra
Coimbra, Portugal
jpvilela@dei.uc.pt

***Abstract*—We consider privacy of obfuscated location reports that can be correlated through time/space to estimate the real position of a user. We propose a user-centric Location Privacy-Preserving Mechanism (LPPM) that protects users not only against single reports, but also over time, against continuous reports. Our proposed mechanism, designated clustering geo-indistinguishability, creates obfuscation clusters to aggregate nearby locations into a single obfuscated location. To evaluate the utility of the mechanism, we resorted to a real use-case based on geofencing. Our evaluation results have shown a suitable privacy-utility trade-off for the proposed clustering geo-indistinguishability mechanism.**

***Index Terms*—Location Privacy, Location Privacy-Preserving Mechanisms, Location-Based Services, Geo-Indistinguishability, Clustering, Geofencing**

I. Introduction

Location privacy has became an emerging topic due to the pervasiveness of Location-Based Services (LBSs). When sharing location, our personal information is exposed to possibly untrustworthy entities which have the capacity to share the collected data with third parties. Although the analysis of this data may be beneficial to several services and, hence, to consumers, the collected data may contain sensitive and private information, thus raising privacy concerns. This is specially critical for location data because human mobility traces are highly unique and extremely predictable given that visited locations can reveal the user's identity, habits, addictions or even health conditions [1]–[3].

Obfuscation mechanisms are commonly used to protect users' location [4], [5], in where an obfuscated version of the exact user location is reported instead of the exact location. These techniques are known as user-centric LPPMs, as privacy is preserved for each user independently, and act at collection time, that is before the data is collected and therefore preserve privacy even against the service provider [6]. Based on the classic notion of differential privacy, geo-indistinguishability has been proposed to design LPPMs that limit the amount of information that is disclosed to a potential adversary observing the reports. Geo-indistinguishability guarantees that any two locations within a given radius around the user are statistically indistinguishable. The Planar Laplace (PL) mechanism was the first proposed geo-indistinguishable LPPM [7]. This mechanism obfuscates the exact user location by adding 2-dimensional Laplacian noise centred at the user location.

While promising, geo-indistinguishability considers reports to be independent from each other, thus discarding the potential threat that arises from exploring the correlation between reports. In fact, in the context of sporadic release of data, LPPMs typically consider reports to be independent between each other [8]. However, location data can be reported sporadically or continuously, depending on the LBS [8], [9]. The frequency of reports impacts the achieved privacy level, since the adversaries can use the inherent correlation between reports to improve their attacks [1], [5], [10], [11]. Although some recent research has started considering temporal and spatial correlations [10], [12], this topic is far from being mature and is still considered an open issue [5]. Therefore, we explore this intrinsic characteristic of location data from the perspective of privacy protection by proposing a user-centric LPPM that acts at collection time and that is suitable for continuous reports of location data. The contributions of the work are summarised as follows:

- We develop a new mechanism that takes into consideration the distance between the reported locations and the frequency of updates. The clustering geo-indistinguishability creates obfuscation clusters for closer locations, such that the mechanism returns the same obfuscated point for nearby locations. We evaluate and compare the developed mechanism with the PL mechanism and a recently proposed adaptive version of the PL to the continuous scenario [13]. Results showed that the proposed mechanism provides a better trade-off between privacy and utility.
- We assess the utility of our mechanism through a real use-case based on geofencing. To the best of our knowledge, we are the first to consider a practical geofence application as a utility metric.

The remainder of the paper is organised as follows. Section II presents background concepts and the implemented mechanisms. Section III describes clustering geo-indistinguishability, our proposed LPPM. Section IV presents the evaluation of the proposed mechanism, specifying the experimental setup and the conducted methodology. Finally, Section V draws the final conclusions of this work.

II. Background

Location privacy is an emerging topic of research [1], [4], [5] due to the pervasiveness of LBSs and always connected mobile devices. Shokri *et al.* classified LBSs as continuous or sporadic depending on the frequency of location reports [8]. An LBS is considered continuous when the user's location is reported periodically, and it is considered sporadic when the user requests a single location query, receives the result from the service and then terminates the query.

The existing LPPMs have been developed for both continuous [10], [14] and sporadic scenarios [7], [15], [16], depending on whether they consider the dependence or independence

978-1-7281-0875-9/19/$31.00 ©2019 European Union

of the temporal correlation between subsequent reports, respectively. Earlier research focused on the sporadic scenario, whereas recently, studies on continuous reports have been emerging.

Considering the objective of this work, we selected one mechanism for each scenario. In particular, we focus on the PL [7] for the sporadic scenarios and on the adaptive geo-indistinguishability [13] for the continuous scenarios. The PL mechanism was selected for the sporadic scenario, since it was the first mechanism that achieved the notion of geo-indistinguishability, which provides formal privacy guarantees of differential privacy applied to location data. The adaptive geo-indistinguishability was selected for the continuous scenario, since it is a recent mechanism based on the PL that explores the correlation between reports for protecting location privacy. The following subsections detail these LPPMs, respectively.

A. Geo-Indistinguishability

The geo-indistinguishable PL consists of adding 2-dimensional Laplacian noise centred at the exact user location x and with the following Laplacian distribution, whose probability density function (pdf) is:

$$p(z|x) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d_{\mathrm{x}}(x,z)} \tag{1}$$

To obtain z from x using equation (1), we can add a randomly drawn vector expressed as a radius r and angle Θ. In this case, Θ is uniformly chosen from $[0, 2\pi)$ and r is computed by drawing p uniformly from $[0, 1)$ and feeding it to the inverse planar Laplacian cumulative distribution function. This function is calculated using the negative branch W_{-1} of the Lambert W function and is defined as $C^{-1}(p) = -\frac{1}{\epsilon}\left(W_{-1}\left(\frac{p-1}{e}\right)+1\right)$. Therefore, the obfuscation location z is calculated by $z = x + \langle r\cos\Theta, r\sin\Theta\rangle$.

B. Adaptive Geo-Indistinguishability

The adaptive geo-indistinguishability was proposed for continuous scenarios. This mechanism uses the PL with a dynamic ϵ that is computed according to the correlation between the new location and the past locations. Based on this correlation, the adaptive mechanism adjusts the amount of noise required to obfuscate the exact user location x. Thus, the mechanism increases the privacy level when the correlation between reports is high and improves the utility level when the correlation between reports is low. The correlation is measured as the error between an estimation and the exact user location, where the estimation is obtained using a simple linear regression. Formally, we can define the dynamic ϵ as follows [17]:

$$\epsilon = \begin{cases} \alpha \times \epsilon, & \text{if } d(x,\hat{x}) < \Delta_1 \\ \epsilon, & \text{if } \Delta_1 \leq d(x,\hat{x}) < \Delta_2 \\ \beta \times \epsilon, & \text{if } d(x,\hat{x}) \geq \Delta_2 \end{cases} \tag{2}$$

where x is the exact user location, $\hat{x}$ is the estimation, $d(\cdot)$ is the euclidean distance, Δ_1 and Δ_2 are two thresholds, and α and β are two constants. The authors also specify the following constraints: $\Delta_2 > \Delta_1$, $0 < \alpha < 1$, and $\beta > 1$. In original work [13], the authors used the following parameters: $\Delta_1 = 0.96/\epsilon$, $\Delta_2 = 2.7/\epsilon$, $\alpha = 0.1$, and $\beta = 5$. From the first branch of the equation (2), we have that if the distance between the exact user location and the estimation is lower than a small threshold Δ_1, i.e. high correlation, then privacy should be improved. To do so, ϵ is decreased by a factor $\alpha < 1$. On the other hand, when the error is larger than a higher threshold Δ_2, i.e. low correlation, the utility is enhanced by multiplying ϵ with the factor $\beta > 1$ (third branch). Otherwise, when the error is between $[\Delta_1, \Delta_2[$, the value of ϵ does not change.

III. Clustering Geo-Indistinguishability

In order to develop a new mechanism that can be used both in the sporadic scenario and in the continuous scenario, we started by looking at the PL, which is considered the state-of-the-art LPPM for the sporadic scenario. From the PL mechanism, we know that the exact user location x is reported as an obfuscated location z, which is obtained by adding 2-dimensional Laplacian noise centred at the exact user location. Since the frequency of updates was shown to have impact on the privacy preservation of the user location [11], our idea consists in creating a mechanism that obfuscates the exact user location x by applying the PL mechanism; then, an obfuscation cluster centred at the real location x is created, such that the same obfuscated location z is reported for every real location inside the cluster. With this approach, we take advantage of the original PL for sporadic scenarios (i.e. low sampling frequency and distant reports), while providing a solution that leads to the same obfuscated report for continuous scenarios, in which real locations are close by.

Our mechanism produces an obfuscated location z_i for the first user location x_i, by directly applying the PL mechanism. The location x_i creates an obfuscation cluster centred at $x_c = x_i$, that is, a circle centred at x_c, whose obfuscated point is z_i. For the next user location x_{i+1}, the mechanism verifies if it is inside the area of the previous obfuscation cluster centred at x_c. If the user location is inside the area, the mechanism reports the previous obfuscated point, that is, $z_{i+1} = z_i$. Otherwise, the LPPM obfuscates the location x_{i+1} with the PL mechanism and creates a new obfuscation cluster centred at $x_c = x_{i+1}$. In order to verify if the user location is inside the area of the previous cluster, the mechanism calculates the distance d between the current location and the location that originated the previous cluster x_c, using the great circle distance $g(\cdot)$. The parameters of our scheme are then the radius of the obfuscation cluster and the value of the privacy parameter ϵ. To reduce the number of parameters, which in turn increases the usability of the mechanism, one can set r to depend on the ϵ value, according to the original definition of PL, such that $\epsilon = l/r$.

Algorithm 1 shows the implemented approach. The parameters of the algorithm are the exact user location x_i, the privacy parameter ϵ and the radius of obfuscation r. By applying this algorithm, the user location x_i will be obfuscated and the algorithm will return the obfuscated location z_i. Regarding the parameters, the value of ϵ will be used to apply the PL mechanism and the radius r will be used to compute the radius of the obfuscation area of the clusters as explained above.

A. Privacy Analysis

The correlation between reports may degrade the privacy level of the LPPMs [11]. In particular, when a user reports several nearby points, the PL mechanism leads to the disclosure of user information [7]. For instance, if we consider the most continuous scenario possible, i.e. when the user is continuously reporting the same location, the PL mechanism will produce

Algorithm 1 LPPM based on clustering

```
function CLUSTERING(x_i, ε, r)
    if first report then
        x_c = x_i
        z_i = planarLaplace(x_i, ε)
    else:
        distance = g(x_c, x_i)
        if distance ≤ r then
            z_i = z_{i-1}
        else
            x_c = x_i
            z_i = planarLaplace(x_i, ε)
    return z_i
```

several obfuscated locations for that same user location. From the obfuscated locations and considering the behaviour of the Laplacian distribution used by the mechanism, an adversary can delineate the centre of the obfuscation area, which enables to discover the exact user location. Our proposed mechanism prevents this situation, since the clustering mechanism reports the same obfuscated point for nearby locations, which is a clear advantage.

Regarding the nearby locations, by the definition of ϵ-geo-indistinguishability, if the distance between two locations x, x' is at most r, then the multiplicative distance between the obfuscation pdf centred at x and x' is at most l, where l is the level of privacy. Thus, for closer locations, the distributions are similar and, consequently, the probability of generating the same obfuscated location is higher. Since our mechanism uses the same obfuscated point for locations that dist at most r, it is guaranteed that the obfuscated locations reported by the proposed mechanism are geo-indistinguishable, thus avoiding the need to use PL to produce a new obfuscated location.

Furthermore, it can be shown that the privacy level of geo-indistinguishability scales linearly with the number of queries [7]. That is, for n location queries applying geo-indistinguishability independently to each query results in $n\epsilon$ privacy disclosure. On the other hand, each time our proposed mechanism uses the previous obfuscated location, it avoids a new application of the PL to produce a different obfuscated location. Therefore, our mechanism prevents the linear privacy degradation of geo-indistinguishability that comes from multiple applications of the protection mechanism.

Lastly, although the number of reported points does not decrease, as a result of applying our mechanism, the reported point does not change in some of the cases. In fact, once the user reports the same location instead of a new location, for the service, the user stays in the same location. Thus, there is less disclosure of user's information. This is specially relevant for the continuous scenario where the frequency of updates is higher and, consequently, the distance between reports is smaller. Our mechanism takes advantage of this property and thus greatly reduces the number of applied obfuscations to nearby locations.

IV. Evaluation

The following subsections describe the experimental setup, the performance of the proposed clustering geo-indistinguishability mechanism and evaluate the achieved levels of privacy and utility. Moreover, we present the comparison between our mechanism and two existing mechanisms, the PL and the adaptive geo-indistinguishability.

A. Experimental Setup

To evaluate the effectiveness of the proposed LPPM, we selected an attack mechanism and a real mobility dataset. Since adversaries may use maps to locate the users [4], we selected a state-of-the-art Map-Matching (MM) technique [18], which enables us to locate vehicles on road networks. While MM is usually applied for Global Positioning System (GPS) navigation, it can also be employed as a mechanism for tracking attacks against location privacy [11], particularly in the continuous scenario.

This subsection describes the experimental setup for the evaluation, namely, the attack mechanism, the used metrics, the selected dataset and its pre-processing.

1) Map-Matching: The objective of MM is to find a path that corresponds to a sequence of location reports, assuming that these reports are noisy and follow a normal distribution. To do so, [19] resorts to a road network and a Hidden Markov Model (HMM), where the HMM's hidden states at each noisy location correspond to potential locations on the road. The most likely path from the HMM is obtained using a Viterbi algorithm

MM can be used as a pre-processing technique in an LBS, where the location reports are mapped to the most likely position for the exact location. Nevertheless, MM can also be used by an adversary to track a user even if the latter is using an LPPM, since an LPPM acts as a noisy channel.

2) Metrics: To measure the privacy level, we can use a point-by-point metric, such as the adversary error. The adversary error measures the correctness of an adversary through the distance between the exact user locations and the adversary's estimations. The adversary estimation error is computed as $P_{AE} = E\{d(x_i, \hat{x}_i)\}$, where the adversary error (AE) is the expected distance $d(\cdot)$ between the exact user location x_i and the adversary's estimation $\hat{x}_i$. Typically, Euclidean distance is used as distance metric [20].

However, for a tracking attack, a point-by-point metric would fail to assess the effectiveness of the tracking mechanism. The authors of [19] define $F\text{-}score$, also called F_1 score, to evaluate the accuracy of the MM, which can be calculated by the following equations.

$$\begin{aligned} precision &= \frac{L_{correct}}{L_{matched}}; \quad recall = \frac{L_{correct}}{L_{truth}} \\ F_1 score &= 2 \times \frac{precision \times recall}{precision + recall} \end{aligned} \tag{3}$$

where $L_{matched}$ is the length of the output path, L_{truth} is the length of the corresponding ground-truth and $L_{correct}$ is the length of the portions of the output path that overlap with the ground-truth path. This metric basically measures how accurate the mechanism is through the amount of overlapped path, $L_{correct}$, between the adversary's estimated path, $L_{matched}$, and the ground-truth's path, L_{truth}. The value of F_1 score varies between 0 (worst path match) and 1 (best path match).

3) Dataset Selection: We have selected the *Taxi Cabs in USA* dataset [21], that contains vehicular trajectories with a high sampling rate, thus being appropriate for analysis of continuous scenarios, as well as allowing sub-sampling to mimic different frequencies of updates. This dataset contains

trajectories from over 500 taxis travelling in the area of San Francisco with duration of 30 days. This data includes not only geo-location collected through a GPS at an average rate of 10 seconds, but also the occupancy of the taxi.

4) Dataset Pre-Processing: We started by selecting a set of relevant trajectories as follows. We first limited the distribution of trajectories to a bounding box over the peninsula of San Francisco, as this is the most dense area, defined from south and west by the coordinates (37.5996104427, -122.5168704724) and from north and east by the coordinates (37.81093499, -122.3535056708). Then we considered only trajectories with passengers, where the flag of occupancy is true [22]. This division allowed us to remove cases where the taxi was stopped waiting for a client. Finally, we selected trajectories with a duration of at least 1 hour, with intervals between reports of at most 100 seconds, to avoid temporal discontinuities between reports. This pre-processing resulted in 46 trajectories. To observe if the dataset contained noisy readings, we displayed the trajectories in the map and did a manual inspection of some of these trajectories, which confirmed our premise. For example, there were some GPS locations reported in the ocean instead of in the bridge that the vehicle was clearly crossing.

To enhance the original data, we first apply the MM mechanism described in Section IV-A1 to the 46 trajectories from the original dataset, by employing the same parameters as in [23], which is the baseline to the work in [19] and uses GPS data as in our case. In [23] the estimated standard deviation was $\sigma = 6.86$m and they limited the potential locations to a bounding box of 50m centred in the noisy GPS reading o_i. For the other parameters, we use the original values of [19]: $\lambda_y = 0.69$ and $\lambda_z = 13.35$. The constraint of the 50m radius around o_i produced observations without candidate points due to the existing nodes of the road network. For these observations, we consider the nearest node of the road network as candidate. Moreover, after further manual inspection, we observed that in some of the trajectories the taxi stays roughly in the same place, which we attribute to heavy traffic. Consequently, we removed those trajectories and we obtained 30 trajectories as test data, henceforth referred as our ground-truth.

Finally, to vary the frequency of reports we subsample the dataset by suppressing reports such that the interval between consecutive points is at least Δ_t. Since our focus was on continuous scenarios, we selected the following set of values: $\Delta_t = [60, 120, 180, 240, 300, 360, 420, 480, 540, 600]$ seconds. It should be noticed that the values in our set are already considered low-sampling rate in the context of MM [18], [24]. In the selected MM technique [19], they consider a range of frequencies between 60 and 300 seconds.

B. Methodology

Figure 1 summarises the employed methodology. As explained in Section IV-A4, the GPS data is pre-processed using the MM technique, resulting in the ground-truth, which in turn is subsampled considering the aforementioned values of Δ_t. Then the LPPMs are applied to the subsampled data, i.e. to the exact locations. Finally, the MM is executed on the obfuscated locations to obtain the adversary's estimations. To evaluate the privacy level of the LPPMs, we used the average adversary error, P_{AE}, as a point-by-point metric, and the F_1 score from equation (3) as a trajectory metric. Moreover, to evaluate the trade-off between the privacy and the utility of the LPPMs, we resort to a real use-case based on geofencing.

1) LPPMs Configuration: In this work, LPPMs are applied and evaluated under multiple values of ϵ as follows $\epsilon = [0.016, 0.032, 0.064, 0.128]$ m^{-1}. As parameters of the adaptive geo-indistinguishability mechanism, we started by using the values defined in the original work. However, we observed that the adaptive geo-indistinguishability was benefiting the utility instead of the privacy level for the majority of the values of ϵ and Δ_t, since most instances appeared above the Δ_2 threshold (third branch of equation (2)). In order to have diversity in the behaviour of the adaptive mechanism, we selected two different values of Δ_1 and Δ_2. Figure 2 shows the boxplot of the estimation errors with the selected thresholds, $\Delta_1 = 750$ and $\Delta_2 = 1750$. This ensures that we encompass a set of scenarios in which adaptive geo-indistinguishability optimises for privacy (lower values of Δ_t, where most instances are below Δ_1), utility (higher values of Δ_t) and intermediate cases. Lastly, regarding the simple linear regression, we chose to use the parrot function because it exhibited the best results [17]. The parrot function simply consists of returning the previous value as the prediction.

Regarding the selection of the obfuscation radius r of the clustering mechanism, we resorted to the original definition of PL, such that $\epsilon = l/r$ or, likewise, $r = l/\epsilon$. For that, we considered the above set of ϵ values and $l = \log(4)$, as suggested by the authors [7]. As we can observe, the same value of r can be obtained with different combinations of values of ϵ and l. Therefore, the degrees of freedom of our mechanism actually correspond to the value of ϵ and r, with r being a function of ϵ. As such, we focus our analysis on the effect of ϵ.

2) MM Configuration: The parameters σ, λ_y and λ_z for the MM attack were estimated following the proposal of the authors in [19]. To estimate the σ, we calculated the standard deviation of the location measurement errors. To estimate λ_y and λ_z, we measured the circuitousness and the temporal implausibility for a selected group of trajectories. Regarding the selection of the trajectories, the authors used the paths with duration between 1 and 5 minutes, resulting in 4828 trajectories with an average length of 2.6 km. In the same way, we used the trajectories with duration between 1 and 5 minutes that had at least 2 km of travelled distance, resulting in 6003 trajectories. The estimation of the parameters resulted in the following values: $\lambda_y \approx 0.07$ and $\lambda_z \approx 0.74$.

Furthermore, considering the efficiency of the attack, we only take into account candidate points within a radius computed for MM. To compute this radius, we use the inverse distribution function of the Gaussian distribution, such that the circle centred at the observation contains the exact location with 90% probability. Intuitively, this corresponds to the case where the attacker computes the set of potential locations, where with 90% probability the exact location is in the set. When there is not a candidate within this radius, we consider the nearest node of the road network as candidate. The road network used covers the area defined by the referred bounding box and was obtained from OpenStreetMap using the OSMnx tool [25]. The road network is in the form of a *networkx multidigraph*, which is manipulated using the NetworkX tool [26].

C. Number of Points per Cluster

Figure 3 shows the average number of points per cluster obtained by applying clustering geo-indistinguishability as

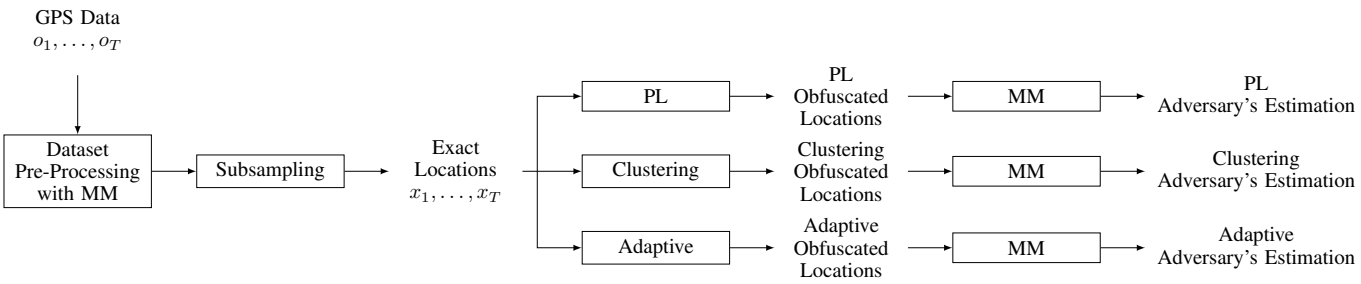


Fig. 1: Diagram of the followed methodology.

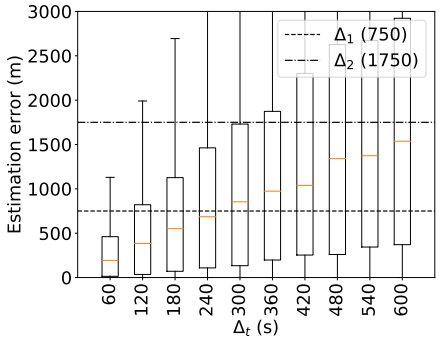


Fig. 2: Boxplot of the estimation errors of the adaptive geo-indistinguishability with varying minimum interval between points Δ_t. Dashed lines correspond to the thresholds Δ_1 and Δ_2.

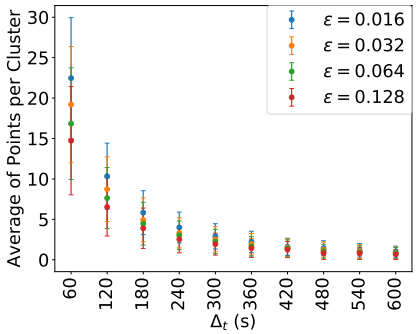


Fig. 3: Average of points per cluster obtained by applying the clustering geo-indistinguishability for different values of ϵ, with varying minimum interval between points Δ_t, and respective 95% confidence intervals.

a function of Δ_t for various ϵ values. As aforementioned, the clusters were created according to the obfuscation radius, that is related to the value of ϵ. Thus, for the set of ϵ values, the used set of radiuses is approximately $r = [86.64, 43.32, 21.66, 10.83]$ m. As expected, for lower values of Δ_t, the number of points per cluster is higher than for higher values of Δ_t. This can be explained by the proximity of the locations when the time interval is lower. In particular, from this figure, we can observe that the number of points per cluster is approximately less than three for the values of $\Delta_t \geq 360$ s in all values of ϵ that we used. The other five values of Δ_t correspond to time intervals between 60 to 300 seconds, that is, from the case where the user is reporting at every minute until the case where the user is reporting every 5 minutes. Therefore, as we will detail in the following subsections, the impact of our mechanism on privacy and utility will be higher for values of $\Delta_t \leq 300$ s, that is, for time intervals smaller or equal than 5 minutes.

Furthermore, we can observe from Figure 3 that the average number of points per cluster increases with the decrease of the ϵ value, for all Δ_t values. This can be explained by the original definition of PL, such that $\epsilon = l/r$ and, likewise, $r = l/\epsilon$. From this definition, we have that a higher value of ϵ corresponds to a smaller radius r. Thus, the radius of the obfuscation clusters is smaller for higher values of ϵ and, consequently, there are less points per cluster for those ϵ values.

D. Privacy Evaluation

To evaluate the privacy of the mechanism, we used the adversary error as a point-by-point metric, and the F_1 score as a trajectory metric. The obtained results will be compared with the results of the PL and the adaptive geo-indistinguishability.

1) Adversary Error Metric: Figure 4 presents the average adversary error of the three mechanisms as a function of Δ_t for various ϵ values. As we can observe, the results of the clustering geo-indistinguishability are similar to the results of the PL mechanism. In fact, these results reveal that our mechanism maintains the privacy level point-by-point.

When we compare the results of the adaptive geo-indistinguishability with the results of the clustering geo-indistinguishability, we observe that the difference between the average adversary error is less than $\sim$10 m for $\epsilon = [0.016, 0.032]$ and $\Delta_t \geq 420$ s, for $\epsilon = 0.064$ and $\Delta_t \geq 300$ s, and for $\epsilon = 0.128$ and $\Delta_t \geq 240$ s. For the remaining cases, the adaptive geo-indistinguishability has a bigger adversary error, which can be explained by the fact that the adaptive mechanism is mostly benefiting the privacy level in those cases. As mentioned in Section II-B, this behaviour is a consequence of the parameters used in the adaptive mechanism. From equation (2) and from Figure 2, we have that: $\epsilon = \beta \times \epsilon$ when the estimation errors are greater than Δ_2; $\epsilon = \epsilon$ when the estimation errors are between Δ_1 and Δ_2; and $\epsilon = \alpha \times \epsilon$ when the estimation errors are lower than Δ_1. Therefore, as we can observe in Figure 2, the majority of the estimation errors for values of $\Delta_t \leq 240$ s is lower than Δ_1, then the mechanism improves the privacy level by increasing the obfuscation level, which results in larger adversary errors. For the remaining values of Δ_t, the mechanism does not change the value of ϵ or improves the utility, by increasing the value of ϵ, which results in lower values of adversary error.

2) F_1 Score Metric: Figure 5 shows the comparison between the three mechanisms. When we compare clustering

geo-indistinguishability with the PL mechanism, we can observe that the clustering mechanism has lower values of F_1 score for all values of $\Delta_t < 360$ s and all values of ϵ, which means higher privacy level. For the remaining values of Δ_t, F_1 score is lower in some values of Δ_t and ϵ and slightly higher in others. As we showed in Section IV-C, the number of points per cluster is higher for $\Delta_t < 360$ s and, therefore, the impact of our mechanism is more significant for these values of Δ_t.

Regarding the adaptive geo-indistinguishability, we can observe that the results of the F_1 score become similar to the results of the clustering geo-indistinguishability with the increase of the Δ_t values, which can be explained by the behaviour of the adaptive mechanism. Since for higher values of Δ_t, the estimation errors are higher and, consequently, the mechanism tends to improve the utility of the data. Moreover, the difference between the F_1 score of the adaptive and the clustering mechanisms is less than $\sim 5\%$ for $\epsilon = [0.016, 0.032]$ and $\Delta_t \geq 300$ s, and for $\epsilon = [0.064, 0.128]$ and $\Delta_t \geq 240$ s. From Figure 5, we can further observe that the clustering geo-indistinguishability has lower values of F_1 score in some of these cases.

Lastly, we can observe from Figure 5 that for lower values of Δ_t and ϵ, the adaptive mechanism has an F_1 score of about 20%. Recalling the meaning of this metric, this value translates to an overlap between the output path and the ground-truth path of approximately 20%. Thus, the mechanism discloses less than a quarter of the original trajectory. While this is advantageous from a privacy perspective, it leads to a severe degradation of utility as we will now illustrate.

E. Utility Evaluation

To evaluate the utility of the mechanisms, we consider a real use-case based on geofencing. Geofencing is the process of generating virtual geographical perimeters/areas in where events occur when users enter or leave such perimeters. A location service provider can create geofences around locations of interest (e.g. Points of Interest (PoIs)), such that users traversing the geofence can receive relevant information with respect to the location (e.g. marketing or discounts from supermarkets).

Therefore, we created geofences to several PoIs from San Francisco. In order to have diversity of PoIs, we used PoIs from different domains, namely: hotels, museums and supermarkets. These PoIs were obtained from OpenStreetMap using the OSMnx tool [25], resulting in a total of 524 PoIs. Moreover, the geofences were created under multiple values of radius r. The used set was defined as $r = [100, 200, 300, 500, 1000]$ m. This set was chosen according to the guidelines for creating geofences for android developers [27], where a minimum radius of 100-150 m is recommended.

In our work, when a user enters in the area of a geofence, the application retrieves this PoI. Thus, we executed the application for the ground-truth user mobility locations to obtain the ground-truth PoIs. Then, we executed the application for the obfuscated user mobility locations that result from applying the PL, the adaptive geo-indistinguishability and the clustering geo-indistinguishability, to obtain the reported PoIs. Finally, in order to measure how the reported PoIs match the ground-truth PoIs, we used the classification true/false positive/negative.

To classify the results as True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN), we first defined the positive class and the negative class. Since the objective of the application is to return PoIs, we define returning a PoI as the positive class and returning *None* as the negative class. When the reported PoI is equal to the ground-truth PoI, we have a TP. When both the ground-truth and the reported do not return any PoI, we have a TN. When the ground-truth returns a PoI or *None* and the reported returns a different PoI, we have an FP. Lastly, when the reported returns *None* and the ground-truth returned a PoI, we have an FN. This classification is summarised in Table I.

TABLE I: Classification True/False Positive/Negative.

Ground-Truth	Reported	Classification
A given PoI	Correct PoI	True Positive
None	None	True Negative
A given PoI, None	Incorrect PoI	False Positive
A given PoI	None	False Negative

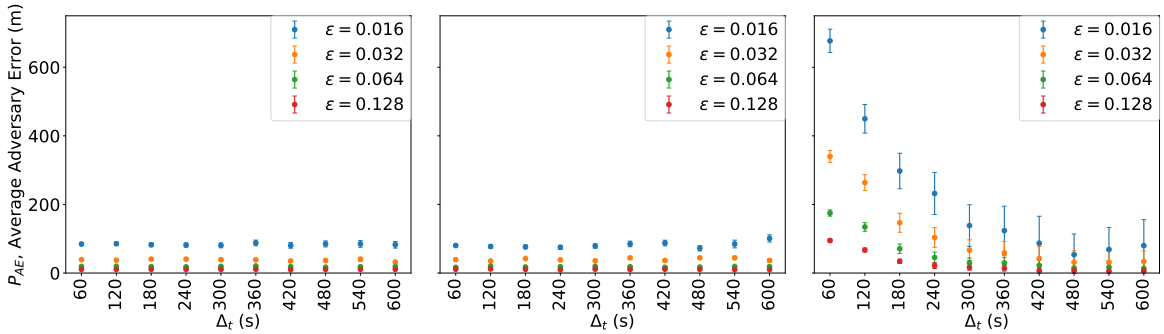
Based on this classification, we were interested in knowing how many PoIs were correctly or incorrectly identified. To do so, we used the True Positive Rate (TPR) and the False Positive Rate (FPR), which are defined as follows:

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN} \tag{4}$$

Although we used the TPR and FPR, we could have used the True Negative Rate (TNR) and the False Negative Rate (FNR) because the metrics are complementary. However, from the point of view of the utility, the TPR is more relevant since it corresponds to the cases of both the ground-truth and the LPPM returning the same PoI.

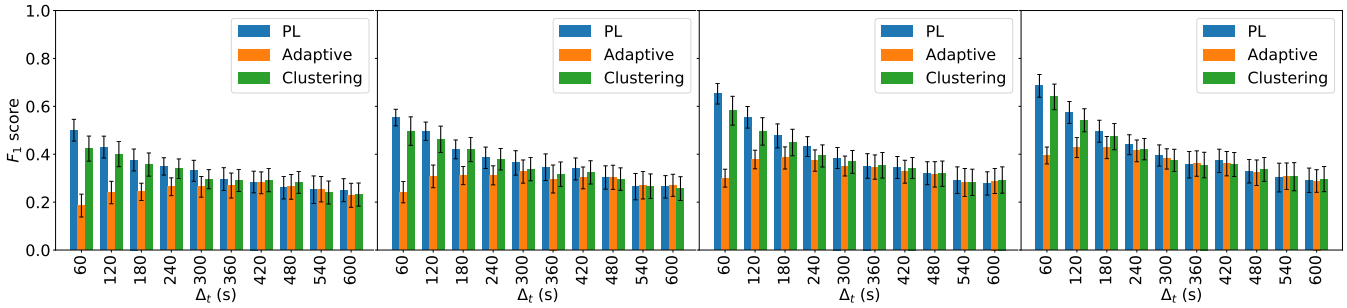
Figures 6a and 6b respectively represent the TPR and the FPR of the three mechanisms averaged for all values of Δ_t, for each ϵ, and for each geofence radius From Figure 6a, we can observe that the TPR of all three mechanisms improves for growing ϵ values. This is expected since higher ϵ values correspond to lower obfuscation and, therefore, obfuscated locations that are closer to the real ones. This effect of ϵ fades away with increasing geofence radius, since a larger radius increases the size of the geofence region and, consequently, benefits the probability of getting the correct PoI, irrespectively of the level of obfuscation applied.

Regarding the comparison between the mechanisms, we can observe that the adaptive geo-indistinguishability has the lowest TPR for all values of ϵ and all values of the geofence radius. As we observed before, the adaptive mechanism has higher adversary errors, which means a higher distance between the reported point and the exact user location. Thus, these results reveal that the adaptive mechanism is improving the privacy level by degrading the utility of the data. On the other hand, the clustering geo-indistinguishability has the highest TPR, except for the radius of the geofence 100 m and $\epsilon = 0.016$. This exception can be explained because the $\epsilon = 0.016$ corresponds to an obfuscation radius of approximately 86 m. Thus, as the mechanism creates obfuscation clusters within a radius of 86 m, the distance between the obfuscated locations and the exact user locations included in the cluster can be higher than the radius of the geofence and, hence, the mechanism reports an incorrect PoI. As we mentioned before, when the geofence radius increases, the difference between the TPR of the clustering geo-indistinguishability and the other mechanisms decreases. In particular, when the radius of the geofence is 1 km, the TPR of the three mechanisms is similar for high values of ϵ, since the increase of the radius of the



(a) PL (b) Clustering (c) Adaptive

Fig. 4: Average adversary error and respective 95% confidence intervals of PL, clustering and adaptive mechanisms for different values of geo-indistinguishability privacy parameter ϵ, with varying minimum interval between points Δ_t.



(a) ϵ = 0.016 (b) ϵ = 0.032 (c) ϵ = 0.064 (d) ϵ = 0.128

Fig. 5: Comparison between the F_1 score value of the PL, the adaptive and the clustering geo-indistinguishability for different values of ϵ, with varying minimum interval between points Δ_t, and respective 95% confidence intervals.

geofence, i.e. the increase of the geofence region, benefits the probability of reporting correct PoIs.

Figure 6b shows the FPR of the three mechanisms. Inversely to the TPR, here the FPR decays with increasing ϵ, since higher ϵ values correspond to less obfuscation and, therefore, improved FPR. As we can observe, the adaptive mechanism has the highest value for all geofence radiuses, which means that this mechanism reports more incorrect PoIs. On the other hand, the PL mechanism and the clustering geo-indistinguishability report fewer incorrect PoIs, again with our scheme closely following PL. Lastly, when the radius of the geofence grows, the size of the geofence region increases and, consequently, the probability of reporting PoIs is higher. However, this also leads to a high probability of reporting incorrect PoIs, which explains the increase of the FPR for larger geofence radius.

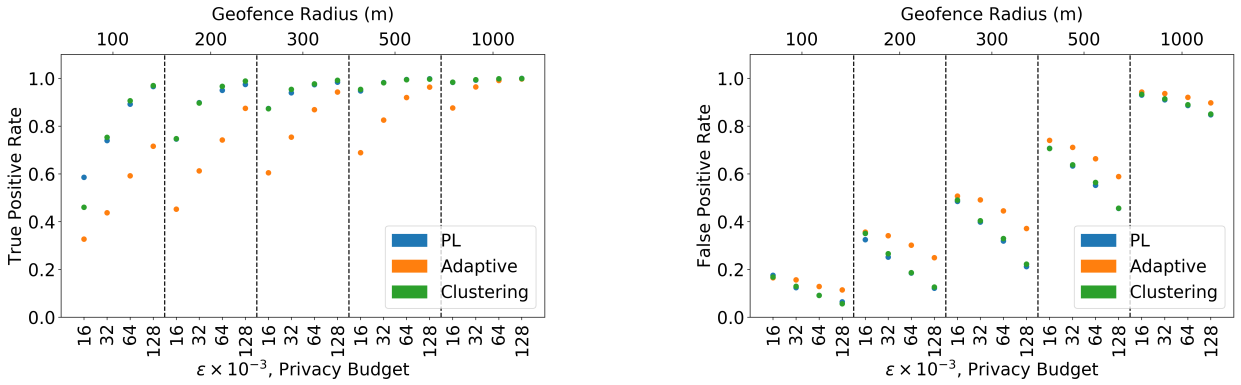
F. Trade-off Between Privacy and Utility

According to the performed evaluation of both the privacy and the utility level of the mechanisms, we can conclude how the mechanisms deal with the trade-off between privacy and utility. In comparison with the PL mechanism, the clustering geo-indistinguishability improves the privacy level for continuous reports of location data (i.e. lower values of Δ_t), with little to no penalty in terms of utility loss (measured by TPR), except for the case of the combined lowest ϵ and lowest geofence radius explained earlier. The comparison of our clustering scheme with adaptive geo-indistinguishability shows that the adaptive mechanism is able to achieve higher privacy levels (i.e. lower F_1 scores) for continuous scenarios (smaller Δ_t values), albeit at a severe cost in terms of utility, as shown in the practical geofence analysis. Therefore, we can conclude that the clustering geo-indistinguishability provides a favourable trade-off between privacy and utility for continuous reports.

V. Conclusion

Location privacy is an emerging topic of research due to the pervasiveness of LBSs. Regardless of the benefits that these services offer to users, the shared data are not always and only used for the initial purpose. In order to protect the users, LPPMs have been proposed. Our objective was to develop a mechanism that protects users not only against single reports but also over time, against continuous reports. Toward this goal, we developed a new mechanism that is suitable for continuous reports of location data and that improves the level of privacy for continuous reports, with limited or no loss in terms of utility.

To develop the mechanism, we took into consideration the geo-temporal correlations, namely the distance between the reported locations and the frequency of updates. Thus, we created a clustering geo-indistinguishability mechanism that creates obfuscation clusters, such that the the same obfuscated point is reported for nearby locations. According to the performed analysis, our mechanism improves the privacy level in comparison with the PL mechanism, with little to no loss in terms of utility. Moreover, although the adaptive geo-indistinguishability exhibits higher privacy levels, it does so



(a) TPR

(b) FPR

Fig. 6: Comparison between the TPR and the FPR of the PL, the adaptive geo-indistinguishability and the clustering geo-indistinguishability for the average of the Δ_t values and for different values of geofence radius and ϵ.

at the cost of an undesirably high loss of utility, as shown by our analysis of a practical geofence application.

Acknowledgment

This work was carried out in the scope of projects SWING2 (PTDC/EEI-TEL/3684/2014) and MobiWise (P2020 SAICT-PAC/001/2015), funded by Fundos Europeus Estruturais e de Investimento (FEEI) through Programa Operacional Competitividade e Internacionalização - COMPETE 2020, by National Funds from FCT - Fundação para a Ciência e a Tecnologia, through project POCI-01-0145-FEDER-016753, and European Union's ERDF (European Regional Development Fund). Ricardo Mendes wishes to acknowledge the Portuguese funding institution FCT - Foundation for Science and Technology for supporting his research under the Ph.D. grant SFRH/BD/128599/2017.

References

[1] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, 2009.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[3] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Nature Scientific reports*, vol. 3, p. 1376, 2013.

[4] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Personal and ubiquitous computing*, vol. 18, no. 1, pp. 163–175, 2014.

[5] B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang, "Location privacy and its applications: A systematic study," *IEEE Access*, vol. 6, pp. 17606–17624, 2018.

[6] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, June 2017.

[7] M. Andrés, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *20th ACM Conference on Computer and Communications Security*, pp. 901–914, ACM, 2013.

[8] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying location privacy: the case of sporadic location exposure," in *International Symposium on Privacy Enhancing Technologies Symposium*, pp. 57–76, Springer, 2011.

[9] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *2011 IEEE symposium on security and privacy*, pp. 247–262, IEEE, 2011.

[10] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1298–1309, ACM, 2015.

[11] R. Mendes, M. Cunha, and J. P. Vilela, "Impact of frequency of location reports on the privacy level of geo-indistinguishability," (submitted for publication).

[12] H. Liu, X. Li, H. Li, J. Ma, and X. Ma, "Spatiotemporal correlation-aware dummy-based privacy protection scheme for location-based services," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pp. 1–9, IEEE, 2017.

[13] R. Al-Dhubhani and J. M. Cazalas, "An adaptive geo-indistinguishability mechanism for continuous LBS queries," *Wireless Networks*, vol. 24, no. 8, pp. 3221–3239, 2018.

[14] R. Shokri, G. Theodorakopoulos, and C. Troncoso, "Privacy games along location traces: A game-theoretic framework for optimizing location privacy," *ACM Transactions on Privacy and Security (TOPS)*, vol. 19, no. 4, p. 11, 2017.

[15] K. Chatzikokolakis, E. Elsalamouny, and C. Palamidessi, "Efficient utility improvement for location privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, pp. 308–328, 2017.

[16] S. Oya, C. Troncoso, and F. Pérez-González, "A tabula rasa approach to sporadic location privacy," *arXiv preprint arXiv:1809.04415*, 2018.

[17] R. Mendes and J. Vilela, "On the effect of update frequency on geo-indistinguishability of mobility traces," in *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pp. 271–276, ACM, 2018.

[18] M. Kubicka, A. Cela, H. Mounier, and S.-I. Niculescu, "Comparative study and application-oriented classification of vehicular map-matching methods," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 2, pp. 150–166, 2018.

[19] G. R. Jagadeesh and T. Srikanthan, "Online map-matching of noisy and sparse location data with hidden markov and route choice models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2423–2434, 2017.

[20] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: optimal strategy against localization attacks," in *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 617–627, ACM, 2012.

[21] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset epfl/mobility (v. 2009-02-24)." Downloaded from https://crawdad.org/epfl/mobility/20090224, Feb. 2009. (consulted in January 2019).

[22] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi gps traces to social and community dynamics: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 17, 2013.

[23] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet, "Online map-matching based on hidden markov model for real-time traffic sensing applications," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, pp. 776–781, IEEE, 2012.

[24] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp. 336–343, ACM, 2009.

[25] G. Boeing, "Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[26] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[27] Android Developers, "Create and monitor geofences." https://developer.android.com/training/location/geofencing.html. [Online; Accessed in June 2019].