HMC: Robust Privacy Protection of Mobility Data against Multiple Re-Identification Attacks

Mohamed Maouche, Sonia Ben Mokhtar, Sara Bouchenak

HMC: Robust Privacy Protection of Mobility Data against Multiple Re-Identification Attacks

MOHAMED MAOUCHE, SONIA BEN MOKHTAR, and SARA BOUCHENAK, Universite de Lyon, CNRS. INSA Lyon, LIRIS, UMR5250, France

With the wide propagation of handheld devices, more and more mobile sensors are being used by end users on a daily basis. Those sensors could be leveraged to gather useful mobility data for city planners, business analysts and researches. However, gathering and exploiting mobility data raises many privacy threats. Sensitive information such as one's home or work place, hobbies, religious beliefs, political or sexual preferences can be inferred from the gathered data. In the last decade, Location Privacy Protection Mechanisms (LPPMs) have been proposed to protect user data privacy. However existing LPPMs fail at effectively protecting the users as most of them reason on local mobility features: micro-mobility (e.g., individual geographical coordinates) while ignoring higher level mobility features, which may allow attackers to discriminate between users. In this paper we propose $\mathcal{HMC}$ the first LPPM that reasons on the overall user mobility abstracted using heat maps. We evaluate $\mathcal{HMC}$ using four real mobility traces and multiple privacy and utility metrics. The results show that with $\mathcal{HMC}$, across all the datasets 87% of mobile users are successfully protected against re-identification attacks, while others LPPMs only achieve a protection ranging from 43% to 79%. By considering only users protected with a high utility, the proportion of users stays high for $\mathcal{HMC}$ with 75%, while for others LPPMs it goes down to proportions between 4% and 43%.

CCS Concepts: • **Security and privacy → Privacy protections**; **Usability in security and privacy**;

Additional Key Words and Phrases: Location Privacy, Protection Mechanism, Re-identification Attack, Mobility Data, Utility

1 INTRODUCTION

With the unprecedented success of handheld devices, the number of available mobile sensors is increasing. This represents a great resource for city planners, businesses and researches. Examples of such applications include crowd-sensing applications for traffic information (e.g., Nericell [32]), health monitoring (e.g., PEIR [33]), social mechanisms (e.g., fMRI [2]) or research dataset gathering campaigns (e.g., APISENSE [20])

However, the gathering, storage and manipulation of increasing volumes of mobility data opens a number of ethical and legal issues as these data are sensitive in nature and may reveal personal information about individuals (e.g., one's home and workplaces, hobbies, religious, political or sexual preferences).

In order to protect users privacy, a number of data protection mechanisms also called Location Privacy Protection Mechanisms (LPPMs), have been proposed in the literature. The role of an LPPM is to apply data transformations to raw mobility data in order to enforce privacy guarantees to the users. These guarantees can be either well known theoretical properties (e.g., differential privacy [13], k-anonymity [39] and its variants [27] [25]) or more practical techniques to hide sensitive information (e.g., Promesse [37]). To reach this objective LPPMs operate on raw mobility data at various levels of granularity. They may act at the level of individual points (e.g., Geo-I adds noise to individual geo-located coordinates [3]); they may act on a set of co-located points (e.g., Promesse removes clusters of points that correspond to user stops [37]); and they may act at the level of a sub-trace (e.g., W4M enforces k-anonymity by forcing k user traces to be co-located inside the same cylinder [1]) However, none of the existing LPPMs reasons on the users' mobility as a whole considering multiple traces over a period of time (macroscopic vision). This limitation opens the door to powerful user re-identification attacks that try to discriminate users by reasoning on their overall mobility (e.g., AP-Attack [28]), against which existing LPPMs do fall short.

Authors' address: Mohamed Maouche, mohamed.maouche@insa-lyon.fr; Sonia Ben Mokhtar, sonia.benmokhtar@insa-lyon.fr; Sara Bouchenak, sara.bouchenak@insa-lyon.fr, Universite de Lyon, CNRS. INSA Lyon, LIRIS, UMR5250, F69622, France.

In this paper, we propose $\mathcal{HMC}$ (for *Heat Map Confusion*), a Location Privacy Protection Mechanism that protects users against re-identification attacks by reasoning on their mobility as a whole, captured using heat maps. Specifically, in order to protect a dataset of user mobility traces, $\mathcal{HMC}$ first extracts user profiles by aggregating the mobility of each user into a single heat map. Then, $\mathcal{HMC}$ alters each user heat map by making it look similar to the heat map of another user. To limit the decrease in data utility, $\mathcal{HMC}$ uses the heat map of the closest user as a basis for performing the alteration. Finally, $\mathcal{HMC}$ transforms back each altered heat map to a set of mobility traces by trying to retain as much as possible the users' original traces unchanged. The result is a protected mobility dataset on which an attacker that runs user re-identification attacks (e.g., [15] [36] [28]) fails in distinguishing between users.

In this paper, the protection against re-identification attacks is not evaluated only with user re-identification rate as in previous works, but with multiple attacks results. This allows to demonstrate that $\mathcal{HMC}$ does not only protect the users against the attack that also uses heat maps to reason on user mobility but also against attacks that use other models (e.g., points of interests [36] or Mobility Markov chains [16]). Furthermore, we also evaluate data utility using multiple metrics that evaluate data distortion or the accuracy of applications.

To evaluate $\mathcal{HMC}$ we relied on four real mobility datasets [35] [41] [24] [8] and compared $\mathcal{HMC}$ with three representative competitors [3] [37] [1]. We also made $\mathcal{HMC}$ as an open source prototype and included scripts to reproduce our experiments (available at https://github.com/mmaouche-insa/HMC). The results show that $\mathcal{HMC}$ successfully decreases the user re-identification rate of all the attacks. Specifically, across all the datasets using $\mathcal{HMC}$, 87% of mobile users are successfully protected against re-identification attacks, while others LPPMs only achieve a protection ranging from 43% to 79%. By considering only users protected with a high utility, the proportion of users stays high for $\mathcal{HMC}$ with 75%, while for others LPPMs it goes down to proportions between 4% and 43%.

In the remaining of this paper, we review the related work in Section 2. The System Model is presented in Section 3, with the adversary model. In the Section 4, we present the design principles of $\mathcal{HMC}$. Experimental evaluation results are presented in the Section 5. And finally, we draw our conclusions and discuss future work in Section 6.

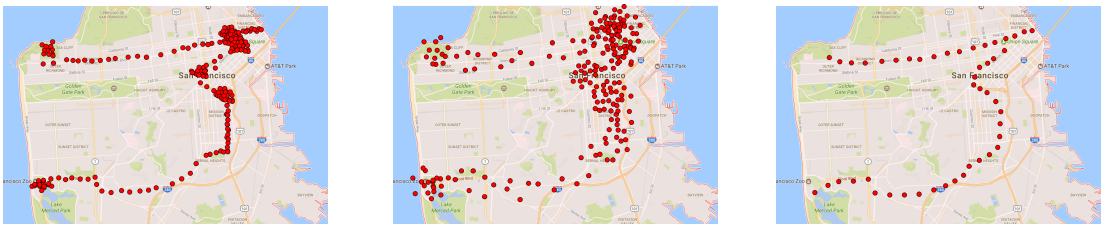
2 RELATED WORK

This section first introduces mobility data, before reviewing the related work in the area of Location Privacy Protection Mechanisms.

2.1 Mobility Data

A mobility data trace is constituted of a sequence of spatio-temporal records $r = (lat, lng, t)$ associated to a given user, where lat and lng correspond to the latitude and longitude of GPS coordinates while t is a time stamp. Figure 1a shows a visual representation of a mobility trace (spacial elements only) of a given user collected in the city of San Fransisco.

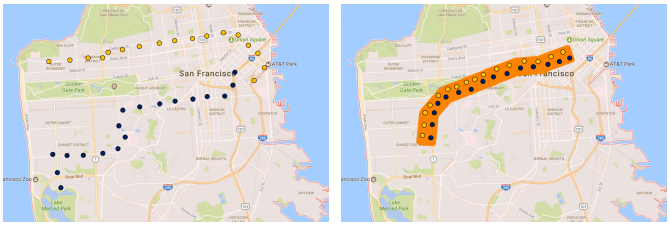
In order to associate semantic information to user raw mobility traces, various mobility models can be built from these traces. An example of such models include the list of users' points of interests (POIs), which are particular places where a user has stopped for a given amount of time. POIs are extracted from raw traces using spatio-temporal clustering algorithms such as [42] [21]. POIs may reveal personal information such as a user's home place, work place or even sexual orientation and religious beliefs if for instance the user often stops at LGBT places or worship places, respectively. This model can be enhanced to a mobility model that includes the probability of moving from one POI to another using Markov chains. This model is richer than the former one as it captures user mobility habits between POIs (e.g., the probability that the user goes to the gym after work). A user's mobility can also be modeled in the form of a heat map in which the intensity of a given cell is related to



(a) One non-obfuscated trace

(b) Example of Geo-I applied to the trace in figure 1a

(c) Example of Promesse applied to the trace in figure 1a



(d) Two non-obfuscated traces

(e) Example of W4M applied to the traces in figure 1d

Fig. 1. Illustration of LPPMs applied to mobility traces

the frequency of user visits in the corresponding area of the map. This representation is a good abstraction of the mobility since in addition to capturing POIs (cells with a high frequency), it takes into consideration higher level features that can discriminate between the users, such as the preferences in paths and a quantification of locations importance. Even though this model does not convey detailed temporal information about the user mobility, it is the only one to capture information about user trajectories.

2.2 Overview of Location Privacy Protection Approaches

To overcome the threats affecting location privacy, Location Privacy Protection Mechanisms (LPPMs) have been proposed in the literature. LPPMs generally take as input a mobility trace (sometimes composed of a single record [23]) or a set of mobility traces and alter these traces in order to produce obfuscated traces. These mechanisms can be applied in two different scenarios. The first one called the *online scenario* is where each gathered record is obfuscated before being sent to the service provider. Example of such LPPMs include *Geo-Indistinguishably* [3], *Mix-zones* techniques (Such as, [5] [14] [34] [26] [11] [12]) where users' ID are switched over in a group of users going through an area called a "mix-zone", for this methods, the mobility data from the mix-zone is not sent in order to hide the ID switch but further work such as Dong and al. [12] used fake traces inside the mix-zone to confuse a possible adversary. Unfortunately, if the traces are gathered to publish a dataset, one real single trace would have multiple IDs and would be considered as coming for multiple different users, as a consequence, the use of user-centric analysis on a total trace would be compromised. Further examples include generalization based techniques such as, *PrivacyGrid* [4], *CloakDroid* [31] where GPS data is discretized using a grid. Generalization can also be done as in the Android Location Privacy Framework [22] where specific locations are replaced by the closest street, city, postal code and more. Other techniques are based on adding

dummy locations to the real mobility trace. For instance, in *SybilQuery* [40] fake traces are generated and sent to the LBS along with the real ones.

A second scenario called the *offline scenario* (the main focus of this paper) is where a set of mobility traces are gathered by a trusted third party that needs to sanitize this data before sharing it with other parties (e.g., a data analytics application running on an untrusted cloud platform). Examples of such mechanisms include, *Gloves* [19] where mobility traces are merged together in order to form anonymity groups. *W4M* [1] where traces are made closer together inside cylindrical volumes. Qardaji and al. [38] used differentially private grids to partition the data space to generalize the data in order to perform Differentially private count queries. Bindschaedler and al. [7] where they create fake traces that share statistical properties with the real traces in order to replace them. In *Promesse* [37], they use a speed-smoothing algorithm in order to erase POIs and thus erasing sensible information about the users.

Further to their usage scenarios, LPPMs are often classified depending on the privacy guarantees they offer to the users. There exist two major privacy guarantees presented in the literature: k-**anonymity** [39] and **Differential Privacy** [13].

2.3 K-Anonymity-Based Location Privacy Protection Mechanisms

The k-anonymity property states that a user is hidden among a set of $k-1$ other users with similar properties [39]. In the context of mobility data this translates to the ability to hide a given user in a geographical zone (called a cloaking area) where there are at least $k-1$ other users [6]. Among the LPPMs that enforce k-anonymity, CliqueCloak [17] use a trusted third party to compute cloaking areas, PRIVE [18] has the same principle but relies on peer-to-peer communication between users to compute the cloaking areas. These two LPPMs allow the protection of a given geo-located point (i.e., online scenario) but do not consider a mobility trace as a whole. Instead, Wait 4 Me (W4M) [1] allows to enforce k-anonymity on mobility traces by extending k-anonymity to (k, δ)-anonymity. In this context, a user mobility trace will be hidden within $k-1$ traces inside a cylindrical volume of radius δ. Figure 1e shows the application of W4M on the two mobility traces of Figure 1d. From these two figures, we observe that the two traces have been distorted to fit into the same cylindrical zone.

2.4 Location Privacy Protection Mechanisms Based on Differential Privacy

Differential privacy was initially proposed for database systems, it ensures that the result of an aggregate query over a table should not be significantly affected by the presence or absence of one single element of this table [13]. This concept has been adapted to mobility data in an LPPM called Geo-Indistinguishability (Geo-I) [3]. In Geo-I, differential privacy is ensured by adding spatial noise to location data generated using a two dimensional Laplacian distribution. An example of applying Geo-I to a mobility trace of Figure 1a is depicted in Figure 1b. In this figure, we observe that each point in the original trace has been translated due to the added noise. As such, it is more difficult to infer information such as user's POIs.

2.5 Other Approaches

In addition to the above LPPMs, there exist other LPPMs that try to protect user mobility traces by removing significant information from the traces such as users POIs. Among these LPPMs, Promesse [37] reaches this objective by distorting the temporal dimension of the mobility trace. Specifically, Promesse erases user POIs by using a speed smoothing technique, which assures that between each successive points in the obfuscated trace the distance and time difference are the same. An example of applying Promesse to a mobility trace of Figure 1a is depicted in Figure 1c. In this figure, we observe that POIs have been removed yet it is still possible to reason about user trajectories.

There exists also techniques that make use of cryptography. For instance, Dong and al.[10] uses attribute based encryption to share the users' exact location only to a selected end-user and share a generalized position to the others. While, this scheme is useful to select end-users to trust, it does not protect the privacy of the user against all the possible consumer of the data . It would be interesting to share its location without trusting its end-user.

While the above LPPMs offer various theoretical or practical guarantees to protect the privacy of the users, it is difficult to guarantee resilience against powerful re-identification attacks with background knowledge. Works such as in [28] show how re-identification attacks are able to break through the protection of state-of-the-art LPPMs with different theoretical and practical guarantees. Our objective in this paper is to propose a novel LPPM that reason on higher level features of the mobility in order to transform users' mobility to avoid re-identification.

3 SYSTEM MODEL

In this section, we present our system model. Starting with the adversary model (Section 3.1). We then show a short experiment to illustrate and motivate the problem in hand.

3.1 Adversary Model : User Re-identification Attacks

Let $\mathbb{U} = \{U_1, U_2, \ldots, U_n\}$ be the set of users in the system. A background knowledge of all the user in the system has been gathered by the adversary. We assume that for each user U_i there is a mobility trace T_i corresponding to her past mobility. Specifically, the set of all mobility traces known by the adversary is noted $\mathbb{KD} = \{T_1, T_2, \ldots, T_n\}$ (where $\mathbb{KD}$ stands for Known user Data). From each of these traces T_i, the adversary builds a user profile $P_i = \mathcal{P}(T_i)$ that characterizes the user mobility and acts as a fingerprint.

A Re-identification attack $\mathcal{A}$ defined in Equation 1 run by the adversary, tries to re-associate an anonymous trace T' from the Unknown user data $\mathbb{UD}$ to a known user profile.

$$\begin{array}{ccccc} A & : & \mathbb{UD} & \rightarrow & \mathbb{U} \\ & & T' & \mapsto & \mathcal{A}(T', \mathbb{KD}) = U_a \end{array} \tag{1}$$

Upon receiving an anonymous mobility trace T'_j, the adversary builds its profile $\mathcal{P}(T'_j)$ then researches in the background knowledge of profiles $\mathbb{P} = \{P_1, P_2, ..., P_n\}$ the most similar profiles with regard to a distance measure d and assigns its identity to the anonymous trace (See Equation 2).

$$\mathcal{ID}(T'_j) \leftarrow \arg\min_{U_k} d(\mathcal{P}(T'_j), P_k) \tag{2}$$

The effectiveness of an attack can be evaluated by re-identifying a limited set of anonymous traces $\mathbb{UD} = \{UD_1, UD_2, \ldots, UD_m\}$ and computing the user re-identification rate described in Equation 3. This *rate* is computed for one attack.

$$r(\mathcal{A}, \mathbb{KD}, \mathbb{UD}) = \frac{\sum\limits_{UD_i} \begin{cases} 1 & \text{If } \mathcal{A}(UD_i, \mathbb{KD}) = \mathcal{ID}(UD_i) \\ 0 & \text{Else} \end{cases}}{|\mathbb{UD}|} \tag{3}$$

To evaluate an LPPM using the user re-identification rate, instead of applying the attack on non-obfuscated $\mathbb{UD}$, we apply the LPPM on $\mathbb{UD}$ and compute $r(A, \mathbb{KD}, \mathcal{LPPM}(\mathbb{UD}))$.

Multiple re-identification attacks using this scheme have been proposed. These attacks differ in the user profile they use to characterize the user's mobility

3.1.1 POI-Attack. is an attack based on users' POIs [36] The profiles outputted by $\mathcal{P}_{POI}$ is a set of POI constructed using clustering algorithms such as [42] [21]. In Equation 4, we represent how the distance between two sets of POIs (ie., profiles) is computed. Where P and Q are the sets of POIs for each trace and $d_{geo}(P_r, Q_t)$ computes the geographical distance between two POIs P_r and Q_t. For each POI in P (resp. Q), we search for the smallest

geographical distance with a POI in Q (resp. P). The distance between P and Q is the median of all the distances found.

$$d_{POISets}(P,Q) = median\left[\{\min_t\left[d_{geo}(P_r,Q_t)\right]\setminus\forall r\}\bigcup\{\min_r\left[d_{geo}(P_r,Q_t)\right]\setminus\forall t\}\right] \tag{4}$$

3.1.2 PIT-Attack. is an attack based on Mobility Markov Chains [15] referred to as the probabilistic inter-POI transition attack The profiles outputted by $\mathcal{P}_{MMC}$ is a Mobility Markov Chain [16]. Each state of the Markov chain is a POI and the transition probability t_{P_i,P_j} represents the probability to go from POI P_i to POI P_j. The POIs $P = (P_1, P_2, \ldots)$ of each Markov Chain are ordered by the number of records clustered to form the POI (the POIs are formed using clustering methods on the records of the mobility trace. As a consequence, each POI is formed of a different number of records). PIT-Attack uses multiple distances, the most successful one is the one presented in the Equation 5, which is a combination of two distances depending on a threshold parameter δ. The two distances combined are : (1) The stationary distance (Equation 6) which sums the weighted geographical distances between each combination of two POIs if the distance is lower then a parameter d_0. (2) The proximity distance (Equation 7) that after ranking the POIs by their weight in each Markov Chain. It adds scores r_i if two POIs of the rank i are closer than a parameter Δ. The score is halved after each rank $r_i = \frac{1}{2}r_{i-1}$ and r_0 is a parameter.

$$d_{stats-prox} \equiv if(d_{stat} < \gamma)\ d_{stat}\ else\ d_{prox} \tag{5}$$

$$d_{stats}(P,Q) = \sum_{P_i,Q_j \in P\times Q} w(P_i) \times \begin{cases} d_{geo}(P_i,Q_j) & \text{If } d_{geo}(P_i,Q_j) < d_0 \\ 0 & \text{Else} \end{cases} \tag{6}$$

$$d_{prox}(P,Q) = \left(\sum_{i=1}^{min(|P|,|Q|)} \begin{cases} r_i & \text{If } d_{geo}(P_i,Q_i) < \Delta \\ 0 & \text{Else} \end{cases}\right)^{-1} \tag{7}$$

3.1.3 AP-Attack. is a heatmap based attack [28]. The profile outputted by $\mathcal{P}_H$ (noted $\mathcal{H}$ in the remaining of the paper) is a heat map, which is an aggregate representation of the mobility trace T. In order to construct the heat map, we first divide the world map into square cells of size c. Then, for each cell coordinates (i, j), $H(i, j)$ represents *the probability to be in the cell* (i, j) *of size* c *for the owner of the trace* T. To estimate $H = \mathcal{H}(T)$ using T, we count the number of records in each cell divided by the total number of records in T.

To measure the dissimilarity between two heat maps, the Topsoe divergence is used as defined in Equation 8.

$$d_{Topsoe}(P,Q) = \sum_{i,j} P_{i,j} \ln\left[\frac{2P_{i,j}}{P_{i,j}+Q_{i,j}}\right] + \sum_{i,j} Q_{i,j} \ln\left[\frac{2Q_{i,j}}{P_{i,j}+Q_{i,j}}\right] \tag{8}$$

3.2 Problem Illustration

The Figure 2 depicts the results of three re-identification attacks : POI-Attack [36], PIT-Attack [15], AP-Attack [28] performed on a dataset obfuscated using three representative state-of-the-art LPPMs : Geo-I [3], which enforces Differential privacy; Promesse [37], which erases user POIs and W4M [1] which enforce k-anonymity.

From this figure, we observe a high level of re-identification rate for AP-Attack, the attack based on the usage of heatmaps as user profiles. This shows how well the heat map captures the singularity of user mobility.

Furthermore, with an LPPM such as Promesse that focuses on erasing the users POIs, while we notice that it greatly reduces (nullify in the case of this dataset) the re-identification rate for POI-based attacks, it fails at

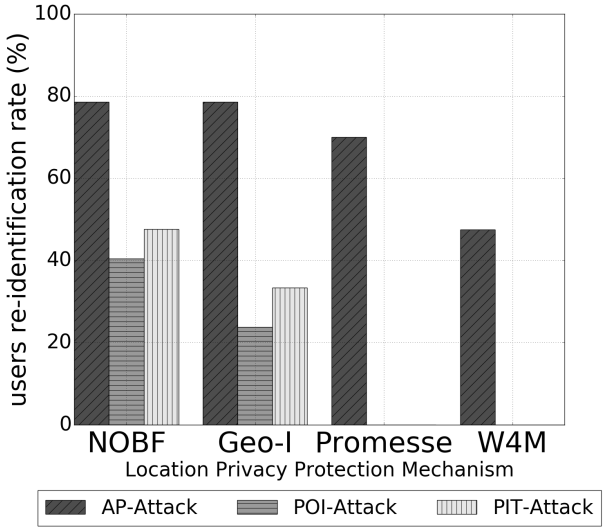


Fig. 2. User re-identification with Geolife mobility data, three state-of-the-art LPPMs, and three re-identification attacks

protecting users against the attack based on heat maps. Lastly, we observe that W4M succeeds at protecting users against POI-based attacks without explicitly erasing POIs, but further studies of the resulting traces have shown that a large portion of the records are erased or moved to enforce the k-anonymity, which raises questions on the future utility of the data (utility evaluated during our experimentations in Section 5.6). We also observe that W4M is not fully capable of protecting the users against an attack based on heat maps. In consequence, we aim at designing an LPPM able to protect users against attacks on heat maps.

4 $\mathcal{HMC}$: HEAT MAP CONFUSION-BASED LOCATION PRIVACY PROTECTION MECHANISM

In this section, we present $\mathcal{HMC}$ a novel LPPM that protects users against re-identification attacks (see Section 3.1). It hides the user by altering its mobility trace using its heat map representation. The heat map is first altered to be the most similar to the heat map of an other user in the background profiles. Then starting from this obfuscated heat map, we construct an obfuscated mobility trace. The goal of $\mathcal{HMC}$ is to deceive the attacker by making the re-identification fall to the wrong identity while maintaining a high utility in terms of map coverage.

Further, we present an overview of how $\mathcal{HMC}$ is constructed, we describe the internal blocks of $\mathcal{HMC}$, the process of Heat Map Alteration and the construction of the outputted mobility trace.

4.1 $\mathcal{HMC}$ Overview

The process of obfuscating a mobility trace T whose identity $\mathcal{ID}(T) = a$ using $\mathcal{HMC}$ is depicted in the Figure 3. This process is composed of three phases :

(1) **Heat Map Creation ($\mathcal{H}$)** : The objective is to construct the heat map of the mobility trace T waiting to be obfuscated. The method is based on the heat map representation of the mobility trace. In consequence, we start by computing $H = \mathcal{H}(T)$ using the heat map Construction module as done in Section 3.1.3.
(2) **Heat Map Alteration (HMA)**: The objective of this phase is to transform H into H', an obfuscated heat map that is more similar to a user profile different than the one of user $\mathcal{ID}(T)$. There is actually more than

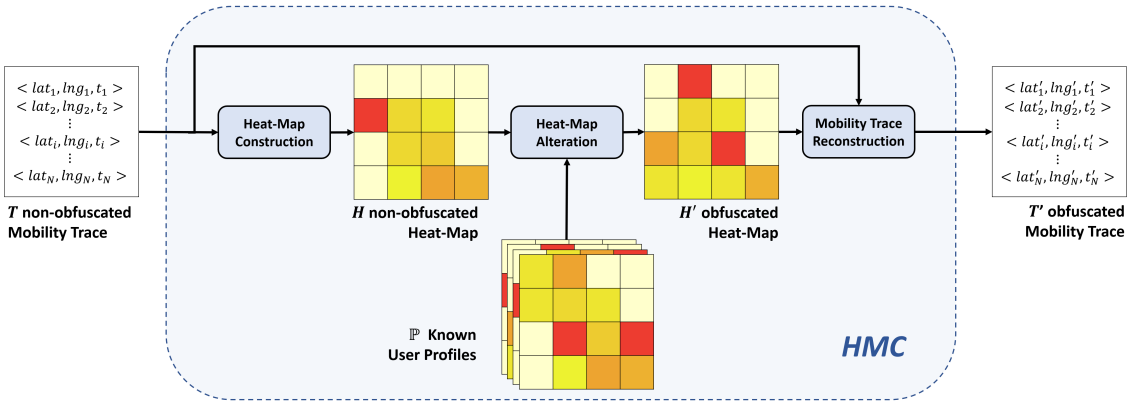


Fig. 3. Overview of HMC

one heat map the satisfies this property (See Equation 9), finding only one is sufficient.

$$\mathcal{HMA}(H, \mathbb{P}) = \{H' \mid \exists K : \mathcal{ID}(K) \neq \mathcal{ID}(H) \wedge \arg\min_{P_i \in \mathbb{P}} d(H', P_i) = K\} \tag{9}$$

(3) **Mobility Trace Reconstruction (MTR)** : We construct an obfuscated mobility trace T' whose heat map is H' the obfuscated heat map of H (Equation 10). We also use T to construct T' in order to keep the trace as similar as possible from the one before obfuscation with privacy guarantees as added value.

$$\mathcal{MTR}(H') = \{T' \mid \mathcal{H}(T') = H'\} \tag{10}$$

In the remain of this section. We describe in more details each phase.

4.2 Heat Map Alteration

We need to construct H' a heat map that satisfies the property of the set $\mathcal{HMA}(H, \mathbb{P})$ defined in Equation 9. We chose to design a method based on iterative modification. As depicted in Figure 4, we first search for U the most similar profile in $\mathbb{P}$ and V the profile with the best utility (area coverage described in 5.3.1) in $\mathbb{P} \setminus \{U\}$.

$$U = \arg\min_{P_i \in \mathbb{P}} d(H, P_i) \tag{11}$$

$$V = \arg\max_{P_i \in \mathbb{P} \setminus \{U\}} \mathcal{UT}(H, P_i) \tag{12}$$

if $\mathcal{ID}(U) \neq \mathcal{ID}(H)$ then H already satisfies the property. This means that the user has a behavior (in the sense of pattern of movements and important locations) that is significantly different from her past mobility and does not need obfuscation (Line 3 of Algorithm 1). On the other hand, if the user is at risk of re-identification, the iterative process starts to research H'.

We first transform the heat map back to a version with the number of records per cell rather than a frequency (Line 7). At each iteration a number of records is added to each cell, depending on the weigh computed using the formula in Equation 13. In order, to affect as little as possible the $\mathcal{UT}$, we alter only cells that are already present in H. Furthermore, we want to reinforce points that are present in both H and V but that are not present in U. More specifically, in Algorithm 1 all the process of $\mathcal{HMA}$ is presented. This processes stops after a number

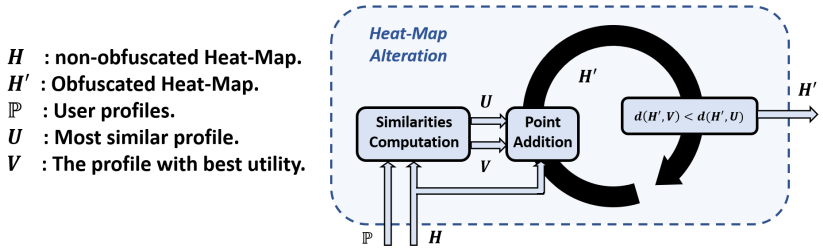


Fig. 4. Heat Map alteration iterative process

Algorithm 1 Algorithm of $\mathcal{HMA}$.

function $\mathcal{HMA}(H,a,\mathbb{P},n,IT_{max})$
  $U \leftarrow \arg\min_{P_i \in \mathbb{P}} d(H, P_i)$ — The most similar profile
  if $\mathcal{ID}(H) \neq \mathcal{ID}(U)$ **then return** H — Does not need obfuscation
  $V \leftarrow \arg\max_{P_i \in \mathbb{P} \setminus \{U\}} \mathcal{AC}(H, P_i)$ — Profile with the best utility
  $c \leftarrow 0$
  while $d(H,V) > d(H,U) \wedge c \leq IT_{max}$ **do**
    $R \leftarrow n \cdot T$
    $W \leftarrow H \odot V \odot (1 - U)$ — $\odot$ represents the pairwise product
    $O \leftarrow R + \left(\frac{a}{\sum W} \cdot W\right)$
    $H' \leftarrow \frac{1}{n} \cdot O$
    $c \leftarrow update(c, H, H', U, V)$ — The counter c rewinds if H' gets closer to V compared to U
    $H \leftarrow H'$
  end while
  if $c = IT_{max}$ **then return** V — If no H' candidate is found, use V
  return H
end function

of iteration without improvement. In this case, V is used as H' since it satisfies the property of Equation 9 at the cost of utility loss.

$$\forall (i,j) : W_{ij} = H_{ij} V_{ij} (1 - U_{ij}) \tag{13}$$

4.3 Mobility Trace Reconstruction

This module generates T' a mobility trace whose aggregation is the heat map H' as expressed by Equation 10. The non-obfuscated mobility trace T is used in order to take into consideration utility metrics such as the spatial distortion $\mathcal{SD}$ (Section 5.3.2). Even though, the abstraction using the heatmaps looses the temporal aspect of the mobility traces, the original mobility trace is used to construct the protected one, in order to keep the temporal aspects as close as possible from the original trace.

We distinguish two types of cells in H' : (1) the cells that are present in T (ie., $H(i,j) \neq 0$) and (2) the cells that are not present in T. For the first case, Figure 5 illustrate how the traces contained inside a cell are altered in order to have the same intensity as the one in the obfuscated heat map H'. To reach this objective, we use a

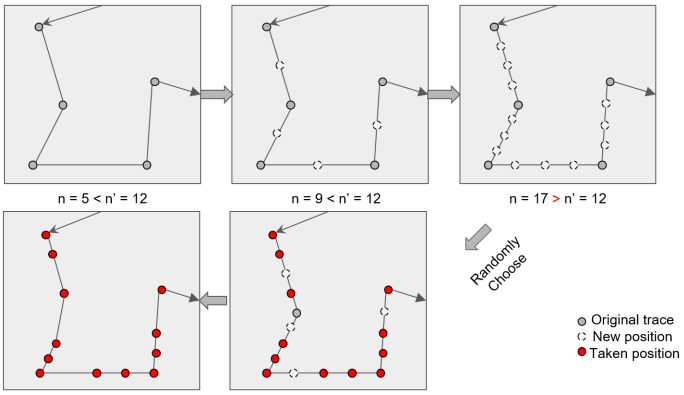


Fig. 5. Cell Number of Records Modification

method inspired from the LPPM Promesse [37]. Specifically, as expressed in Algorithm 2, we use the interpolation between each pair of record in order to create new positions (the timestamp of a new position is equal to the center of the timestamps of the preceding and following record). We do this iteratively until we have enough positions as the number in H' (loop of line 3 to 10 of Algorithm 2). Lastly, we select randomly $H'(i, j)$ records and we keep the timestamps generated during the creation of the positions (line 11 of Algorithm 2). In the cases, where $H'(i, j) < H(i, j)$, we directly select randomly a set of records.

In addition to modify the intensity of the cell, $\mathcal{HMC}$ make sure to not leave small discriminating POIs. That's why, after transforming the number of records in a cell. we make sure to erase small size POIs.

Algorithm 2 Algorithm to adapt the number of records of $\mathbb{R}$ to n records

1: **function** MODIFYNUMBEROFRECORDS($\mathbb{R}$,n)
2: $\mathbb{P} \leftarrow \mathbb{R}$
3: **while** $|\mathbb{P}| < n$ **do** Create new positions in record set $\mathbb{R}$ until its size reaches n
4: $\mathbb{P}' \leftarrow ()$ Empty sequence
5: **for** $i \leftarrow 1$ **to** $|\mathbb{P}|-1$ **do**
6: $p' \leftarrow (p[i-1] + p[i])/2$ Computing the latitude, longitude and timestamp of the middle point
7: $\mathbb{P}' \leftarrow appendToSequence(\mathbb{P}', (p[i-1], p'))$
8: **end for**
9: $\mathbb{P} \leftarrow appendToSequence(\mathbb{P}', (p[|\mathbb{P}|-1]))$
10: **end while**
11: **return** $selectRandomly(\mathbb{P}, n)$
12: **end function**

The second case happens only when no H' is found iteratively and V has to be used as substitute. In this case, we need mobility data in those empty cells in order to apply the interpolation method described above. We use for this a set of records kept from the background knowledge in order to copy real mobility. To be able to put new data, we use time gaps available in the trace (when the GPS is off for instance) to give temporal values to the records. Furthermore, we put a constraint on the portion of trace copied and the temporal gaps using a max speed limit v_{max} as illustrated in Figure 6. We have a trace with a time gap going from the record a to the record

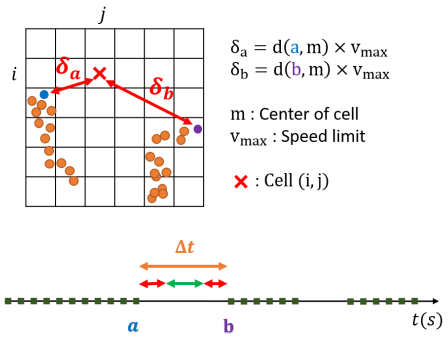


Fig. 6. Time Gaps constraining method

b. Our objective is to generate realistic traces, we ensure that the selected interval is sufficient for a human to move (e.g., at least in walking speed and at most by car) from point a to the cell (i, j) then to b. [1].

In Algorithm 3, we describe how an empty cell is filled with data. It uses as input : $\mathbb{G}$ a list of all available time gaps in the trace, $\mathbb{KD}$ a set of mobility traces to copy mobility from, (i, j) the coordinates of the cell to fill, v_{max} the maximum speed constraining the gaps as explained above, Δt_{max} that limits the time gaps inside a set of records (a set of records from one mobility trace is split into multiple sets of events that respect the limit Δt_{max}), θ_{limits} is the limit of the duration of the set of records to copy (to avoid copying a full day of mobility just to fill one cell).

To fill a cell, we first filter the time gaps according to v_{max} and (i, j) (Line 2) as depicted in Figure 6. Then, we assemble all the data available in the cell (i, j) from $\mathbb{KD}$ after splitting it into multiple sets of records with respect to the constraint Δt_{max}. Next, from all the possible sets of records and all the possible gaps, we select the pair with least distance to connect one another, since a gap has a starting point and an ending point, the distance is the sum of the distance from the start of the gap to the set of records and from the set of records to the ending point of the gap.

4.4 Discussion on Alternatives for $\mathcal{HMC}$

It has to be noted that both the method presented in Section 4.2 and Section 4.3 are pluggable with other methods. The only true conditions for $\mathcal{HMC}$ is to find both H' then T' that satisfies the Equation 9 and Equation 10 respectively. In our instantiation of $\mathcal{HMC}$, we use an iterative method to construct H' and in order to construct T', we use a Time Distortion method [37] and a set of stored mobility traces to avoid using any outsourced library for synthetic mobility trace generation.

Fake user profiles can be used to transform the user behavior to distance her from the behavior of her past self. In this case, the fake profile generated need to be close enough to the user to protect in order to maintain the data utility, but far enough to protect the user identity. With such method, we gain security by avoiding the storage of real user profiles, but we lose the certainty that a user hides from its past self to look like a user that the attacker might re-identify.

[1] A sophisticated trace generator could be used but this is out of the scope of the paper.

Algorithm 3 Algorithm to fill an empty cell (i, j) with real mobility data from $\mathbb{KD}$

function FILLMOBILITYOFCELL($\mathbb{G}$,$\mathbb{KD}$,(i,j),v_{max},Δt_{max},θ_{limit})
  $(\mathbb{G}', p_{index}) \leftarrow filterGaps(\mathbb{G}, v_{max}, (i,j))$
    We keep only the gaps that verify the constrain of speed with respect to v_{max} and (i,j) (see Figure 6)
  $possibleSetsOfEvents = \emptyset$
  for T **in** $\mathbb{KD}$ **with** $getEventsOfCell(T,(i,j))$ **do**
    $splittedSetsOfEvents \leftarrow splitEvents(getEventsOfCell(T,(i,j)), \Delta t_{max})$
    Split the events into multiple sets of events when the time gap between two records exceeds Δt_{max}
    $possibleSetsOfEvents = possibleSetsOfEvents \cup splittedSetsOfEvents$
  end for
  $(setOfEvents, gap) \leftarrow bestMatch(possibleSetsOfEvents, \mathbb{G}, \theta_{limit})$
  $updateGaps(\mathbb{G}, gap)$ Either split the gap or erase it
  $out \leftarrow translateTime(setOfEvents, gap)$
  return out
end function

5 EXPERIMENTAL EVALUATION

In the following, we first present the real-life mobility datasets used in our experiments (Section 5.1). Then, we define the privacy metrics (Section 5.2) and utility metrics (Section 5.3) used in our experiments. In addition, we describe the experimental environment and configuration settings used in the experiments (Section 5.4). Finally, in our experiments, we compare the resilience of $\mathcal{HMC}$ to re-identification attacks with respect to state-of-the-art solutions in Section 5.5 and we further evaluate the utility of the data produced in Section 5.6. Our results show that across all the datasets, $\mathcal{HMC}$ outperforms its competitors in most cases. And for similar privacy results, $\mathcal{HMC}$ has better utility.

5.1 Datasets

We used four real mobility datasets in our experiments. These datasets are: (1) Cabspotting [35] that contains the mobility of 536 cab drivers in the city of San Francisco; (2) Geolife [41] that contains the mobility of 42 users mainly in the city of Beijing; (3) MDC [24] that contains the mobility data of 144 users in the city of Geneva and (4) PrivaMov [8] that contains the mobility of 48 students and staff members in the city of Lyon. A mobility data trace is constituted of a sequence of spatio-temporal records $r = (lat, lng, t)$ associated to a given user, where lat and lng correspond to the latitude and longitude of GPS coordinates while t is a timestamp.

To make the comparison fair between the datasets, we selected in each dataset the 30 most active successive days. We present in the table 1 a description of the datasets used in our experiments. The users are not active in all the days of the period, some are more active than others. We consider as a mobility trace, the mobility of the user during all the period. In all the experiments described in this paper, we split the datasets into a period of 15 days used for the training phase and 15 days used for the obfuscation then re-identification and/or utility evaluation.

Table 1. Description of the datasets

Name	CabSpotting	Geolife	MDC	PrivaMov
# users	536	42	144	48
Localization	San Francisco	Beijing	Geneva	Lyon
# records	11.219.955	1.574.338	904.422	973.684

5.2 Privacy Metrics

In addition to the user re-identification rate presented in Equation 3. we propose to evaluate the anonymity size set of the attacks and to use a multi-attack based privacy evaluation.

5.2.1 k-Anonymity Set Metric: In the adversary model, the attack outputs a single identity. For a user, while not being re-identified fully is a good news (i.e., correct most similar profile). Being the second or third best probable user is still problematic. That's why, we propose this k-anonymity metric. In order to measure for a certain tolerance level k, what is the proportion of user still at risk. This k represents the number of most probable identity for the anonymous trace being re-identified. More formally. The output of a k-attack $\mathcal{A}^{(k)}$ on an anonymous mobility trace T' is a set of k identities with the k most similar profiles. This privacy metric can be seen as a way to measure the k-anonymity set size of a protected mobility trace.

5.2.2 Number of Successful Attacks: This metric computes the number of successful attacks (ie., user correctly re-identified) on a user. It is defined in Equation 14 as a user-centric metric with $\mathbb{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$ being the set of all attacks considered.

$$n(UD, \mathbb{KD}, \mathbb{A}) = \sum_{\mathcal{A}_k \in \mathbb{A}} \begin{cases} 1 & \text{If } \mathcal{A}_k(UD, \mathbb{KD}) = \mathcal{ID}(UD) \\ 0 & \text{Else} \end{cases} \tag{14}$$

Different methods of combining the attacks' results could be used (i.e., $\mathcal{A}' = f(\mathcal{A}_1, \mathcal{A}_2, \ldots)$). For instance, we could leverage the rank results of all the attacks to choose as a result the profile with the best average ranking or use a voting system. Various tests were conducted but the results are inconclusive. Mainly, because AP-Attack is more efficient than the other two attacks and the cases where POI-Attack or PIT-Attack succeed at re-identifying the correct user while AP-Attack fails are rare. In the end, this mix-up of attacks weakens AP-Attack. In consequence, we keep the multi-attack notion by counting the number of successful attacks but we mainly focus on finding the cases where 0 attacks succeeds. This includes the strongest attacks (in our case AP-Attack) but also the cases where POI-Attack or PIT-Attack are the only successful attacks.

5.3 Utility Metrics

The goal of an LPPM is to protect the users' privacy. Unfortunately, the alterations made by the LPPM to the mobility data cause a decrease in the data's utility. Moreover, studies such as [9] make the observation that there is a trade-off between privacy and utility. In consequence, when designing an LPPM, it is important to evaluate the utility of the data produced. Indeed, designing a powerful LPPM that ensure users' privacy without considering the usefulness of data for later analysis is fruitless.

Two approach arise when evaluating the utility of altered data. The first is **data-centric**, which is generic and agnostic of the application. In this case, we consider that every application that is affected by the precision of the data is concerned and could profit from this metric. The second one is **application-centric**. In this case,

we consider a particular application and the conclusion can only be generalized to applications with the same purpose.

In the remaining of this section, we describe the utility metric used accompanied with examples of applications.

5.3.1 Area Coverage: This metric computes how much the alteration affected the regions visited by a user [37]. In other words, while removing records makes places less significant for a user mobility (Ex : Erasing POIs), keeping the information of which regions the user goes through can be important. On the contrary, adding/moving records to new regions adds a fake information that can lead to false deduction for the data analysis. For instance, concluding that a place has many users going through it and in consequence more public transport needs to be available. To compute the Area Coverage $\mathcal{AC}$, the map is dived into equal square regions. For T a mobility trace, $C(T)$ (Eq. 15) returns the set of regions the user goes through, $\mathbb{C}$ represents the set of all possible regions of the dataset and $e \boxdot c$ means that the record e is inside the cell c

$$C(T) = \{c \in \mathbb{C} \mid \exists e \in T : e \boxdot c\} \tag{15}$$

To measure $\mathcal{AC}$ of the obfuscation of T to T', we compute the F-Score value of the precision-recall pair. The precision evaluates the proportion of cells the user goes through in the obfuscated trace which are present in the non-obfuscated trace. While the recall evaluates the proportion of cells of the non-obfuscated trace that are still found in the obfuscated trace.

An example of use case, could be the public health department searching for the areas in the city where the noise disturbance is the most problematic by running a crowd-sensing campaign of noise levels in the city. Precise location are not critic but covering the correct regions of the city is important.

$$\mathcal{AC}_{Precision}(T, T') = \frac{|C(T) \cap C(T')|}{|C(T')|} \tag{16}$$

$$\mathcal{AC}_{Recall}(T, T') = \frac{|C(T) \cap C(T')|}{|C(T)|} \tag{17}$$

$$\mathcal{AC}(T, T') = \mathcal{AC}_{F-Score}(T, T') = \frac{2 \cdot \mathcal{AC}_{Precision}(T, T') \cdot \mathcal{AC}_{Recall}(T, T')}{\mathcal{AC}_{Precision}(T, T') + \mathcal{AC}_{Recall}(T, T')} \tag{18}$$

5.3.2 Spatial Distortion: This parameterless metric computes the spatial error. It considers the traces as polylines $T = (r_1, r_2, \ldots)$ and $T' = (r'_1, r'_2, \ldots)$. For each record x in T' we search for the minimal projection on T. $\mathcal{SD}(\mathcal{T}, \mathcal{T}')$ is the average of the minimal projection of all the records in T'.

$$\mathcal{SD}(T, T') = \frac{1}{|T'|} \sum_{x \in T'} \min_{0<i<|T|} d_{projection}(x, r_i r_{i+1}) \tag{19}$$

An example of use case, could be a city planner wanting to analyze the roads that need the most care by counting the number of users going through them. In this case, a precise spatial location to recognize the correct routes is essential.

5.3.3 Spatio-Temporal Distortion: This metric computes a spatial error constrained by the timestamps of the records. As defined in Equation 21, the spatio-temporal distortion $\mathcal{STD}$ is the average distance between each record of T' and its temporal projection into T. With, the temporal projection of the record $x = (x^{lat}, x^{lon}, x^t)$ in T' being its expected position r_e in T at time x^t. Specifically, we search for $r_i = (r_i^{lat}, r_i^{lon}, r_i^t)$ and $r_{i+1} = (r_{i+1}^{lat}, r_{i+1}^{lon}, r_{i+1}^t)$ in T such as $r_i^t \leq x^t \leq r_{i+1}^t$, then compute r_e the interpolation with the ratio $(x^t - r_i^t)/(r_{i+1}^t - r_i^t)$ (see Equation 20).

An example of use case could be, analyzing users' habits during the day. Such as, which places are mostly visited during the night and need more care in road lights.

$$temporal\_projection(x,T) = \begin{cases} r_1 & \text{If } x^t < r_1^t \\ r_i + \dfrac{x^t - r_i^t}{r_{i+1}^t - r_i^t}(r_{i+1} - r_i) & \text{If } \exists i : r_i^t \leq x^t \leq r_{i+1}^t \\ r_{|T|} & \text{If } x^t > r_{|T|}^t \end{cases} \quad (20)$$

$$\mathcal{STD}(T,T') = \frac{1}{|T'|} \sum_{x \in T'} d_{temporal\_projection}(x,T) \quad (21)$$

5.3.4 Distortion in Surrounding POIs: This metric simulate an application that analysis the POIs surrounding the user location during his mobility. *Open Street Map* [30] is used for this metric. Their open data is uploaded to a MangoDB server and for each record x of the mobility trace in the obfuscated trace T' we query for the surrounding POIs in a rectangular area of size β (with $\mathcal{POI}(x,T,\beta)$). Then, we compare it to the result of the same query for the temporal projection of x in T (see Equation 20) using the harmonic mean of recall/precision (see Equation 22 & 23). The distortion surrounding POIs is the average of all the F-scores of the records of T' (see Equation 24).

$$\mathcal{POI}_{Precision}(x,T,\beta) = \frac{|\mathcal{POI}(temporal\_projection(x,T),\beta) \cap \mathcal{POI}(x,\beta)|}{|\mathcal{POI}(x,\beta))|} \quad (22)$$

$$\mathcal{POI}_{Recall}(x,T,\beta) = \frac{|\mathcal{POI}(temporal\_projection(x,T),\beta) \cap \mathcal{POI}(x,\beta)|}{|\mathcal{POI}(temporal\_projection(x,T),\beta)|} \quad (23)$$

$$\mathcal{DSP}(T,T',\beta) = \frac{1}{|T'|} \sum_{x \in T'} \frac{2 \cdot \mathcal{POI}_{Precision}(x,T,\beta) \cdot \mathcal{POI}_{Recall}(x,T,\beta)}{\mathcal{POI}_{Precision}(x,T,\beta) + \mathcal{POI}_{Recall}(x,T,\beta)} \quad (24)$$

This metric only evaluates if similar POIs are found. It can be extended further to a semantic metric by choosing only certain types of POIs while querying Open Street Map, using the *"amenity"* [29] categorization of the data that references to the type of POI. For instance, one can search for sustenance POIs (i.e., bar, fast food, restaurant, cafe…) or for healthcare POIs (i.e., clinic, dentist, hospital, pharmacy…).

5.3.5 Number of Visits Distortion: This metric simulate a data analysis where the number of visits to a place x is computed for a user. A visit is a record r_i that is within a radius α of x while r_{i-1} is not (See Eq.25). We compute the distortion between the number of visits in the obfuscated trace compared to the non-obfuscated trace (Eq.26).

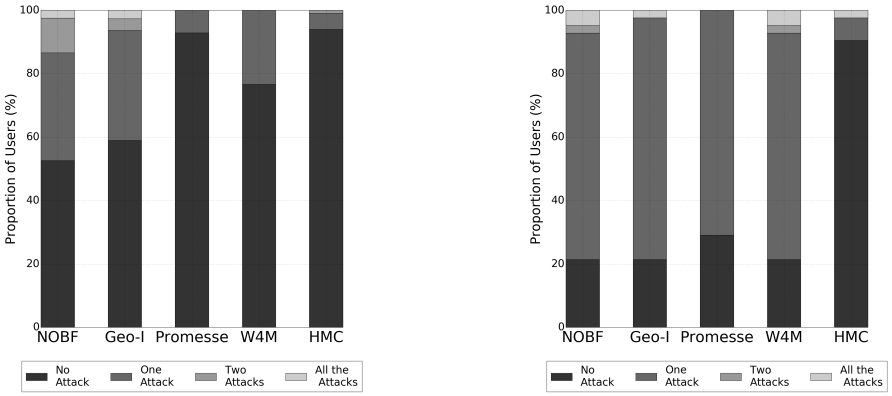
$$\mathcal{NV}(T,x,\alpha) = |\{r_i \in T \mid d(r_i,x) \leq \alpha \wedge\ d(r_{i-1},x) > \alpha \wedge 1 < i \leq |T|\}| \quad (25)$$

$$\mathcal{NVD}(T,T',x,\alpha) = \frac{|\mathcal{NV}(T,x,\alpha) - \mathcal{NV}(T',x,\alpha)|}{\mathcal{NV}(T,x,\alpha)} \quad (26)$$

5.4 Experimental Setup and Configurations

The following experiments were conducted in a computer running an Ubuntu 14.04 OS with 50GB of RAM and 16 cores of 1.2Ghz each. The $\mathcal{HMC}$ prototype is developed in Java & Scala, and runs in the Java Virtual Machine 1.8.0. It is available for download at: https://github.com/mmaouche-insa/HMC

In our experiments, we compare $\mathcal{HMC}$ with three state-of-the-art LPPMs: Geo-I, Promesse and W4M (see Section 2 for more details about these LPPMs). The LPPMs come with their own configuration parameters, that are set as follows. Geo-I's ϵ configuration parameter is set to 0.01; this adds a medium amount of noise to the obfuscated data (the lower ϵ the higher the noise). Promesse's α configuration parameter is set to 200 meters, its represents the distance between two successive sampling points. W4M has two configuration parameters,



(a) CabSpotting (b) GeoLife

Fig. 7. Comparison of $\mathcal{HMC}$ with competitors - Robustness against multiple attacks - CabSpotting & GeoLife datasets

i.e., k that is the minimum number of users inside the cylindrical volume, and δ the radius of the cylindrical volume. Here, k and δ were respectively set to medium values 2 and 600 meters.Finally, $\mathcal{HMC}$'s cell size is set to 800 meters (similar to the good configuration of a heatmap based attack)

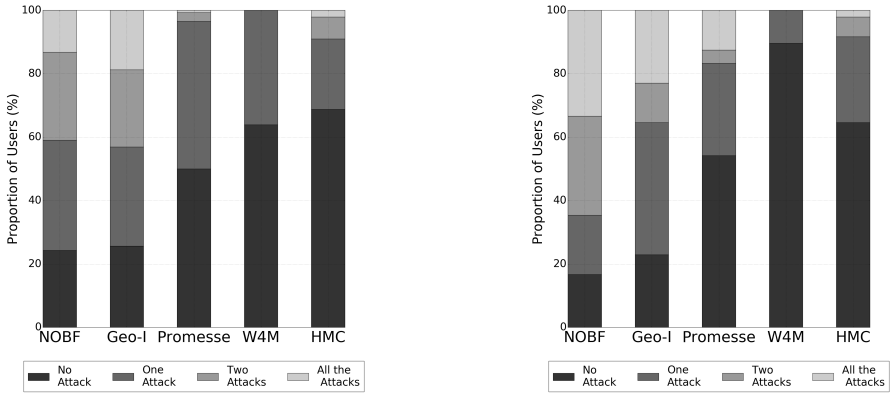
Furthermore, to stress the robustness of the LPPMs and thus evaluating the privacy level they provide, we consider three re-identification attacks in our experiments, namely PIT-Attack, POI-Attack and AP-Attack (described in Section 3.1). The implementations of these attacks have their own configuration parameters. PIT-Attack and POI-Attack have two parameters for the extraction of the POIs from the traces. These parameters are the diameter of the clustering area, and the minimum time spent inside a POI. They were respectively set to 200 meters and 1 hour. And AP-Attack has a configuration parameter that corresponds to the cell size, and that was set to 800 meters. Finally, to evaluate the data utility level provided by the LPPMs, we consider the three utility metrics (described in Section 5.3) that are configured as follows. The Area Coverage utility metric has a configuration parameter that represents the size of a square region, it is set to $800m$ meters. For the metric evaluating the F-score of the surrounding POIs. Its square bounding-box is of distance 200 meters from the record considered. The utility metric that corresponds to the Number of Visits Distortion has one parameter α set to 100 meters, which is the distance threshold α from the place to the record considered as a visit. And the spatial and spatio-temporal distortion utility metrics do not need configuration.

5.5 Privacy Evaluation

In this section, we compare $\mathcal{HMC}$ against three LPPMs using three re-identification attacks.

5.5.1 Resilience against Multiple-Attacks: Lets start with the proposed multi-privacy metric. Indeed, we merge the numerous attacks as presented in Equation 14 of Section 5.2. In the Figures 7 and 8 we present the results by showing the proportion of users with their corresponding number of successful attacks. We notice that $\mathcal{HMC}$ behaves well with a 0 attack protection of 65% to 94% while W4M does 21% to 89% and Promesse 29% to 92%.

If we compare the proportion per dataset (ie., with $proportion_{LPPM}$ versus $proportion_{HMC}$), $\mathcal{HMC}$ has a proportion of users of -16% better than W4M and -22% better than Promesse in average. If we consider all the users (across all the datasets), 87% of the users obfuscated with $\mathcal{HMC}$ have 0 successful attack. While it is 78%



(a) MDC (b) PrivaMov

Fig. 8. Comparison of $\mathcal{HMC}$ with competitors - Robustness against multiple attacks - MDC & PrivaMov datasets

for Promesse, 72% for W4M and 48% for Geo-I. Thus concluding that $\mathcal{HMC}$ outperforms the other competitors. Even if its facing two out three attacks it is not made for (ie., POI-based attacks).

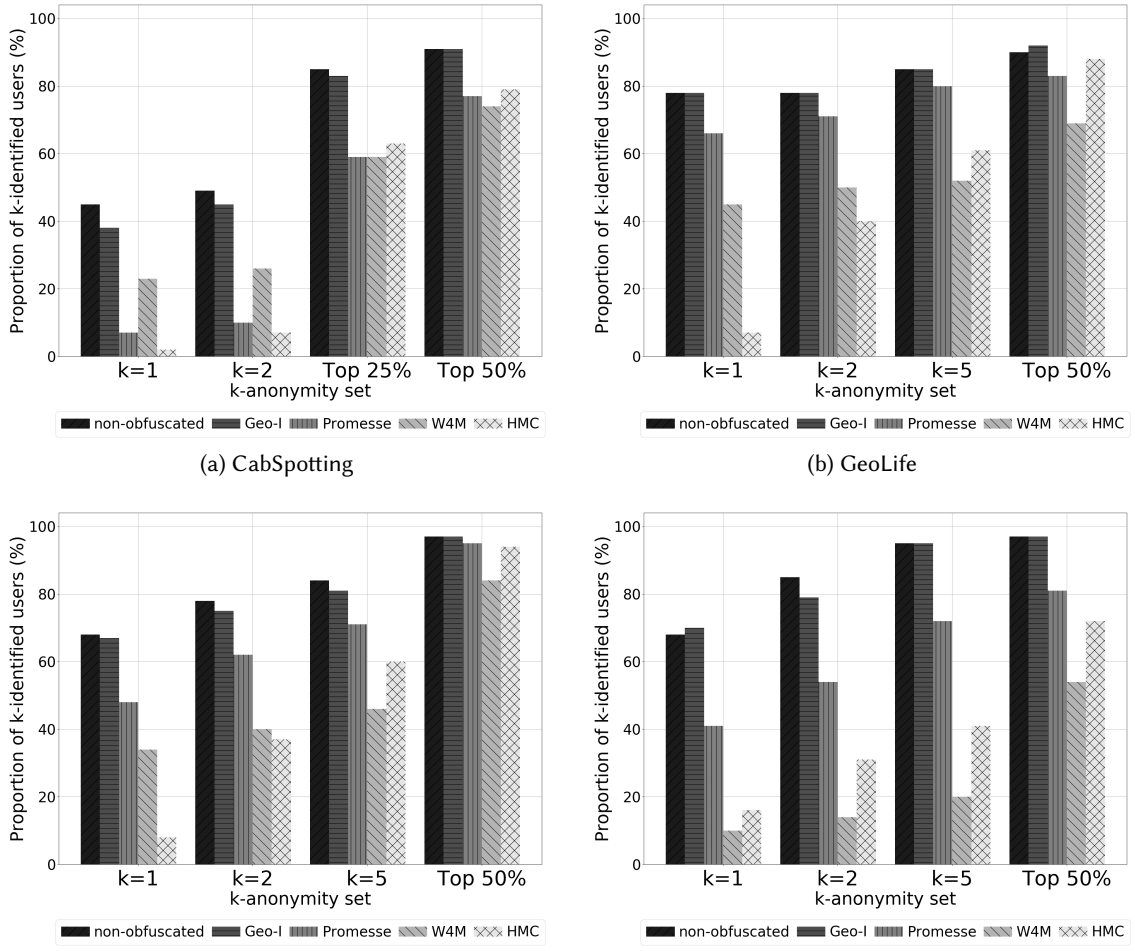
5.5.2 Anonymity Set: For the anonymity set size experiments, the results are depicted in the Figure 16. We notice that for $k = 2$, even though, the gap is getting tighten $\mathcal{HMC}$ still outperform the other LPPMs in three out of the four datasets. The confusion method of $\mathcal{HMC}$ does not transform the mobility to make it the second most similar to its past self but rather to look similar to another user. This why, the correct user does not fall off to the second position but we rather wait for at least $k = 5$. This result come from the fact that the target profile for the confusion is the one with the best utility in area coverage (the confusion is limited to a profile with low utility distortion). Hence, better anonymity size result can be obtained with $\mathcal{HMC}$ by selecting other type of target users (e.g., randomly or k-st most similar profile) but with the cost of lowering the utility. As, it is shown in the next utility experiment, this configuration of $\mathcal{HMC}$ is capable of providing good privacy protection with a better utility than its competitors.

5.5.3 Detailed Resilience against Each Individual Attack: Figures 10, 11 and 12 present the detailed results of user re-identification rate per type of attack, for respectively, AP-Attack, POI-Attack and PIT-Attack introduced in Section 3.1. We first notice that against the strongest attack AP-Attack, $\mathcal{HMC}$ behaves the best. In 3 out 4 of the dataset the rate ranges from 2% to 8% while W4M's rates ranges from 23% to 48%. In the other dataset PrivaMov W4M performs better with 11% over the 19% of $\mathcal{HMC}$. In average $\mathcal{HMC}$ has -20% of user re-identification rate (ie., $r_{W4M} - r_{HMC}$). For POI-Attack and PIT-Attack, $\mathcal{HMC}$ performs worse then W4M but still has low re-identification rates $< 20\%$.

In conclusion, $\mathcal{HMC}$ outperforms the other LPPMs vastly on AP-Attack which was expected since $\mathcal{HMC}$ is based on the heat map representation of the users' mobility. While having good performing results on attacks based on POIs.

5.6 Utility Evaluation

In this Section, we present the utility results of $\mathcal{HMC}$ in area coverage, spatial distortion, spatio-temporal distortion, the distortion in surrounding POIs and the distortion in number of visits.



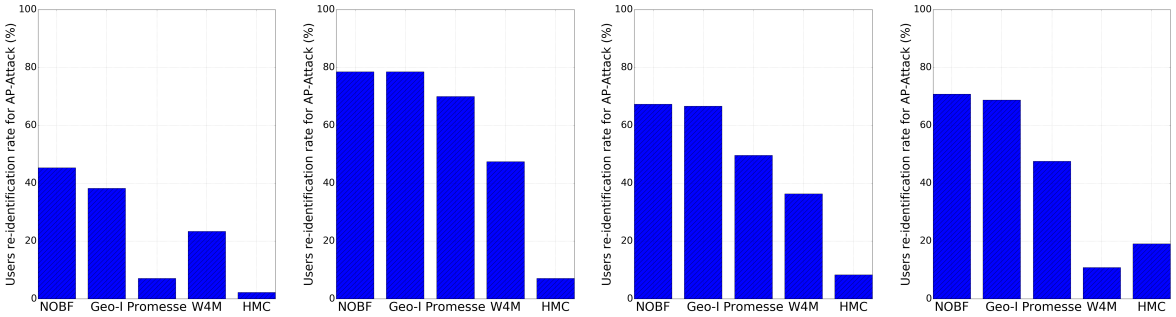
(a) CabSpotting (b) GeoLife

(c) MDC (d) PrivaMov

Fig. 9. Detailed comparison of $\mathcal{HMC}$ with competitors - Anonymity set size against AP-Attack

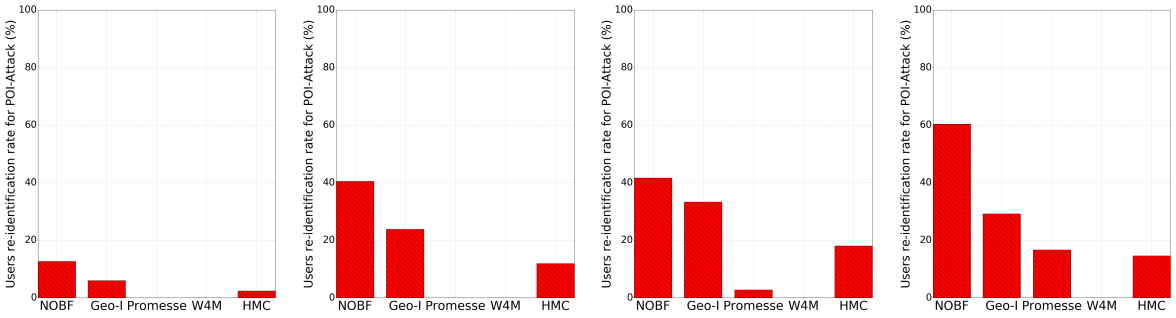
5.6.1 Data-Centric Utility: To present clearly the results, both the metric have a threshold value in which the utility becomes too low for the user. The Table 2 presents those thresholds and the results are depicted in the Figure 13, only the results for the users fully protected by the LPPM are presented (ie., 0 successful attack) because measuring the utility of a non-protected user is insignificant for an LPPM.

We notice that HMC has a big portion of users with High $\mathcal{AC}$ and High $\mathcal{SD}$ it ranges from 27% to 89%, while W4M ranges from 2% to 5% and Promesse 4% to 35%. If we consider all the users across all the dataset, 75% of the user that uses $\mathcal{HMC}$ are fully protected against re-identification attacks and have a high Area Coverage and Spatial Distortion. While it is only 43% for GeoI, 27% for Promesse and as few as 4% for W4M. Overall the only datasets where $\mathcal{HMC}$ is challenged in term of privacy is by over-altering the data and thus lowering the utility. This the case for PrivaMov where W4M has a better privacy but few of the user have a high utility (only 2%). In



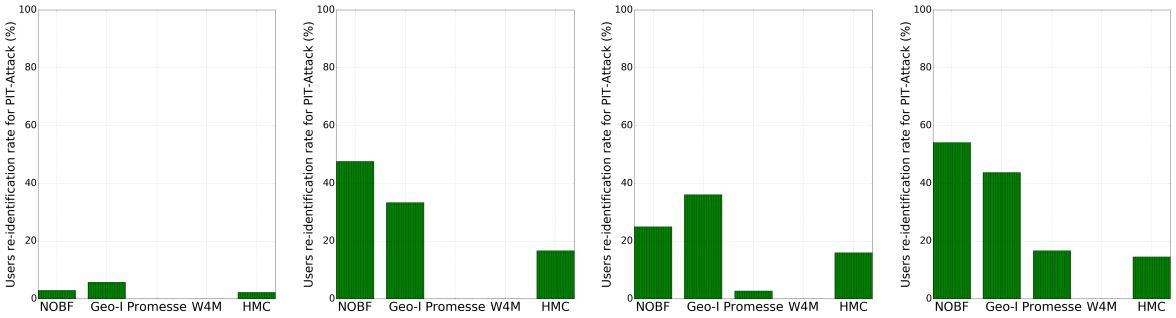
(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 10. Detailed comparison of $\mathcal{HMC}$ with competitors - Robustness against AP-Attack



(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 11. Detailed comparison of $\mathcal{HMC}$ with competitors - Robustness against POI-Attack

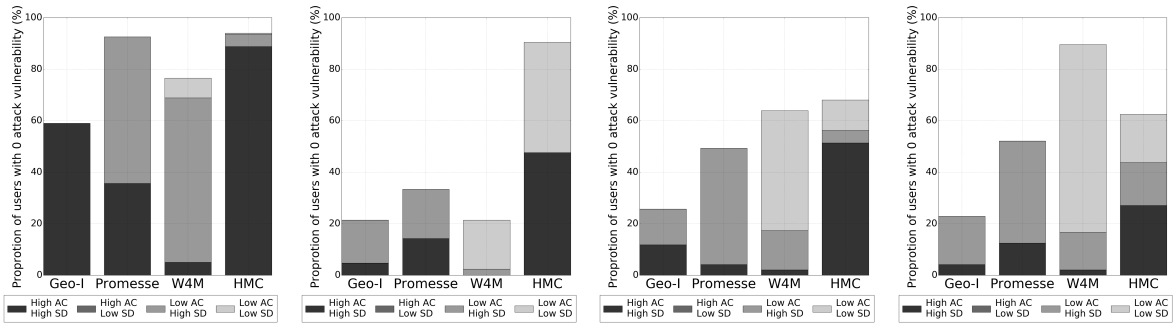


(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 12. Detailed comparison of $\mathcal{HMC}$ with competitors - Robustness against PIT-Attack

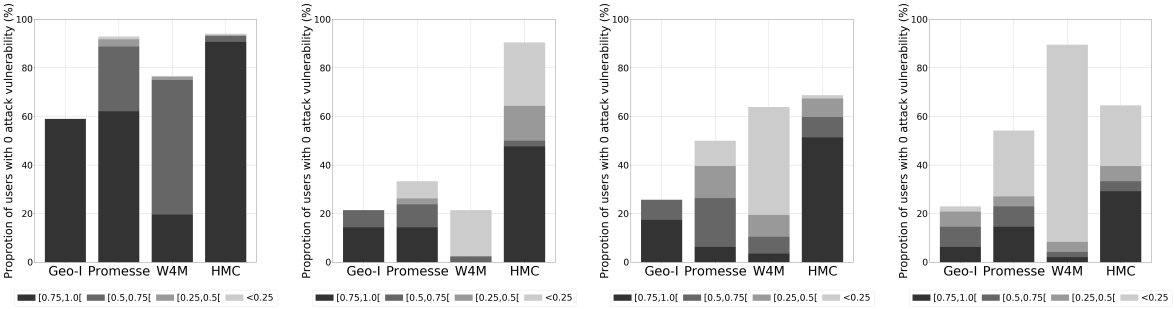
Table 2. Utility Measure Levels Description

	$\mathcal{AC}$	$\mathcal{SD}$
Low	≤ 0.8	$> 200meters$
High	> 0.8	$\leq 200meters$



(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 13. Detailed comparison of $\mathcal{HMC}$ with competitors - Multi-utility metrics evaluation



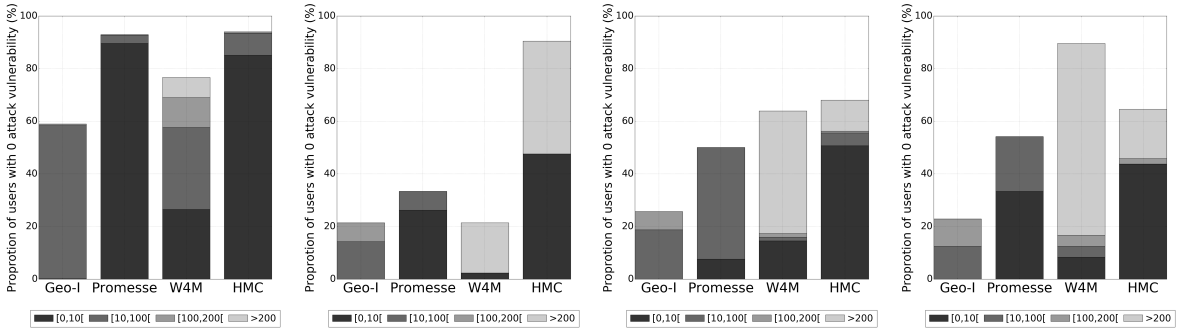
(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 14. Detailed comparison of $\mathcal{HMC}$ with competitors - Area coverage utility metric

Geolife also, Promesse has comparable privacy result but overall half of the users are protected at the cost of lower utility while $\mathcal{HMC}$ protects most of them with high utility.

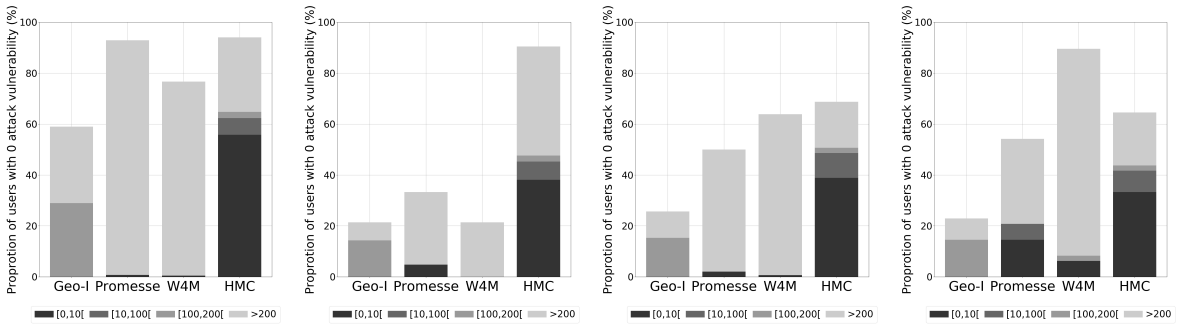
A more detailed analysis for Area Coverage is depicted in Figure 14. We notice that $\mathcal{HMC}$ outperforms all the other LPPMs in term of Area Coverage. $\mathcal{HMC}$'s F-score average ranges from 0.63 to 0.98 while W4M's average ranges from 0.15 to 0.68, for Promesse it is from 0.53 to 0.75 comparable to $\mathcal{HMC}$ but still lower in each dataset.

In term of Spatial Distortion, we present the separate result in the Figure 15. $\mathcal{HMC}$ has medians in centimeters in the datasets. While W4M ranges between $0m$ to $3.6Km$. $\mathcal{HMC}$ has lower values (excluding extreme cases)



(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 15. Detailed comparison of $\mathcal{HMC}$ with competitors - Spatial distortion utility metric



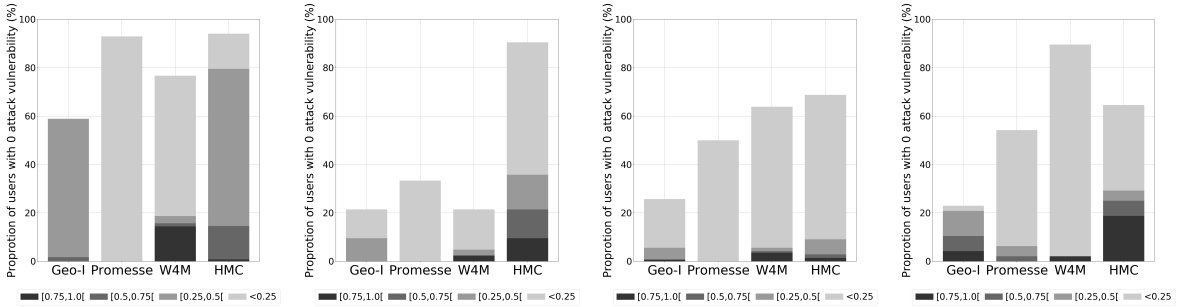
(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 16. Detailed comparison of $\mathcal{HMC}$ with competitors - Spatial-temporal distortion utility metric

thanks to the Promesse-like interpolation technique that create low spatial distortion. That's why Promesse has medians ranging from $4m$ to $13m$.

In term of spatio-temporal distortion, the results are presented in Figure 16. We first notice the results are as expected worse than the spatial distortion. Indeed, the spatio-temporal distortion is the constrained version of the later. Promesse is the LPPM that suffers the most from the temporal constraint, as this method use time distortion in a full portion of trace with speed smoothing in order to erase POIs. Across all the datasets, There is 76% of the users protected with a spatio-temporal distortion greater than 200 meters. While, there is only 27% of users for $\mathcal{HMC}$ (the proportion of users protected and having a spatio-temporal distortion lower than 10 meters is of 50%). W4M already had bad result for the spatial distortion, with a more constraining metric, there is 71% of users across all the datasets that are protected but with a spatio-temporal distortion greater than 200 meters. For Geo-I, even though few of it users are fully protected, there is a systematic noise added to the records, so the there is always a distortion around 200 meters.

5.6.2 Application-centric Utility: We present the result of the comparison of $\mathcal{HMC}$ to the other LPPM with the utility metric that measures the F-score of the query of surrounding POIs (section 5.3.4) in Figure 17. We first



(a) CabSpotting (b) GeoLife (c) MDC (d) PrivaMov

Fig. 17. Detailed comparison of $\mathcal{HMC}$ with competitors - F-scores of surrounding POIs query utility metric

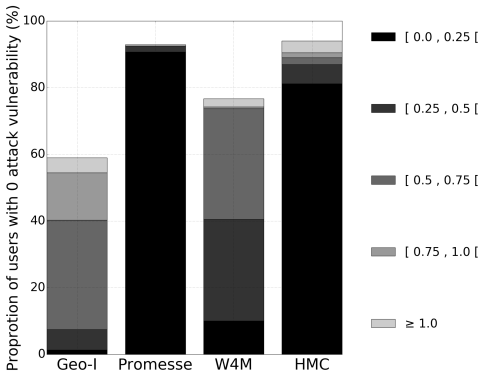


Fig. 18. Comparison of $\mathcal{HMC}$ with competitors - Utility metric in terms of users' number of visits distortion – Cabspotting dataset

notice that with the configuration of 200 meters for the box size, the F-score average is quite low. W4M performs better in Cabspotting for $[0.75, 1]$ interval but in average since $\mathcal{HMC}$ has 64% of users in the $[0.5, 0.75[$ its average F-score is better (0.37 compared to the average F-score of 0.30 of W4M). Except Promese whose average F-scores by dataset ranges from 0.1 to 0.12 the other LPPMs have similar results with a small lead for $\mathcal{HMC}$. Since, $\mathcal{HMC}$ F-scores ranges from 0.13 to 0.39, for W4M it is from 0.13 to 0.30 and Geo-I from 0.11 to 0.42.

For the last utility experiment, we present the result of the visits of "Union Square" in San Francisco (CabSpotting Dataset). We first notice the good results of Promesse by construction with 90% out of the 92% fully protected users have a distortion lower than 0.25. $\mathcal{HMC}$ has similar good results with 81% out of the 94% fully protected user with a distortion lower than 0.25. W4M has a diversity of users with two 30% groups of users with respectively 0.25 to 0.5 distortion and 0.5 to 0.75 distortion, this another low utility level for W4M .

5.7 Discussion

$\mathcal{HMC}$ has good result in utility because it aims at altering the data as few as possible. The cases where new cells of the map are filled are rare and those are the cases where the utility is deteriorated.

We notice that Promesse has lower utility results not because of its perturbation method (which is utility-preserving) but rather because it does't not manage well big time gaps where the user movement was not recorded. Also Promesse and Geo-I apply a systematic perturbation method, even if the user does not need much altering in order to be protected, the utility is always lowered (but still the best to erase POIs). Most importantly while utility-wise, it has good performances, Promesse's poor privacy-results particularly with AP-Attack makes it a bad candidate to protect against the user re-identification threat.

On the other hand, W4M performs poorly utility-wise even in the Cabspotting dataset where numerous users and records are available. Its results on POI-based attacks are good. but far from convincing with AP-Attack. This actually as stated before, the motivation behind the design of the heat map based protection mechanism $\mathcal{HMC}$.

For the case of Geo-I, adding noise deteriorates the utility more than Promesse but it is inept to protect against re-identification attacks, having results similar to non-obfuscated traces. This because of the dependency between successive records. Indeed, this makes the $\epsilon - GeoI$ guarantee looses its power to a $n\epsilon - GeoI$ (n being the number of records).

6 CONCLUSION

In this paper, we presented $\mathcal{HMC}$ a novel LPPM that protects users against re-identification attacks. It uses a heat map alteration process in order to confuse the attacker and to make the re-identification fall to the wrong user. The solution proposed to implement $\mathcal{HMC}$ is based on a iterative modification to transform the heat map and an interpolation technique to alter the number of records in the mobility trace. The heat map is a good abstraction of the mobility as it takes into consideration higher level features that can discriminates between users.

$\mathcal{HMC}$ was evaluated on four real mobility datasets against three representative re-identification attacks and compared to three competitive LPPMs. The evaluation was done using a multi-privacy metric which computes the number of successful re-identification attack and a multi-utility metric with a threshold based Low/High utility categorization simple to interpret. The result show that $\mathcal{HMC}$ outperform the other LPPMs in terms of both privacy and utility.

As future work, the extension of using fake profiles as target user for the confusion would be interesting. And as a direct enhancement, other more sophisticated method to construct fake portion of traces should be incorporated to $\mathcal{HMC}$.

ACKNOWLEDGMENTS

This work benefited from the support of the French National Research Agency (ANR), through the SIBIL-Lab project (ANR-17-LCV2-0014), and the PRIMaTE project (ANR-17-CE25-0017).

REFERENCES

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010.

[2] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mobile Computing*, 7(6):643–659, December 2011.

[3] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. *Ccs'13*, abs/1212.1:–, 2013.

[4] Bhuvan Bamba, Ling Liu, Péter Pesti, and Ting Wang. Supporting anonymous location queries in mobile environments with privacygrid. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, pages 237–246, 2008.

[5] Alastair R. Beresford and Frank Stajano. Mix zones: User privacy in location-aware services. In *2nd IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2004 Workshops), 14-17 March 2004, Orlando, FL, USA*, pages 127–131, 2004.

[6] Claudio Bettini, X Sean Wang, and Sushil Jajodia. Protecting Privacy Against Location-based Personal Identification. In *Proceedings of the Second VDLB International Conference on Secure Data Management*, SDM'05, pages 185–199, Berlin, Heidelberg, 2005. Springer-Verlag.

[7] Vincent Bindschaedler and Reza Shokri. Synthesizing plausible privacy-preserving location traces. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 546–563, 2016.

[8] Antoine Boutet, Sonia Ben Mokhtar, and Vincent Primault. Uniqueness assessment of human mobility on multi-sensor datasets. 2016.

[9] Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Robert Birke, Sara Bouchenak, Lydia Y. Chen, Nicolas Marchand, and Bogdan Robu. PULP: achieving privacy and utility trade-off in user mobility data. In *36th IEEE Symposium on Reliable Distributed Systems, SRDS 2017, Hong Kong, Hong Kong, September 26-29, 2017*, pages 164–173, 2017.

[10] Kai Dong, Tao Gu, XianPing Tao, and Jian Lu. Privacy protection in participatory sensing applications requiring fine-grained locations. In *16th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2010, Shanghai, China, December 8-10, 2010*, pages 9–16, 2010.

[11] Kai Dong, Tao Gu, XianPing Tao, and Jian Lu. Jointcache: Collaborative path confusion through lightweight P2P communication. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM 2013 Workshops, San Diego, CA, USA, March 18-22, 2013*, pages 352–355, 2013.

[12] Kai Dong, Tao Gu, XianPing Tao, and Jian Lv. Complete bipartite anonymity for location privacy. *J. Comput. Sci. Technol.*, 29(6):1094–1110, 2014.

[13] Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[14] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. On the optimal placement of mix zones. In *Privacy Enhancing Technologies, 9th International Symposium, PETS 2009, Seattle, WA, USA, August 5-7, 2009. Proceedings*, pages 216–234, 2009.

[15] Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. De-anonymization Attack on Geolocated Data. *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pages 789–797, 2013.

[16] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Nez Del Prado Cortez. Show Me How You Move and I Will Tell You Who You Are. *Transactions on Data Privacy*, 4:103–126, 2011.

[17] Bugra Gedik and Ling Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *Proceedings of the 25th IEEE International Conference on Distributed Computing Systems*, ICDCS '05, pages 620–629, Washington, DC, USA, 2005. IEEE Computer Society.

[18] Gabriel Ghinita, Panos Kalnis, and Spiros Skiadopoulos. PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 371–380, New York, NY, USA, 2007. ACM.

[19] Marco Gramaglia and Marco Fiore. Hiding Mobile Traffic Fingerprints with GLOVE. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, pages 26:1—-26:13, New York, NY, USA, 2015. ACM.

[20] Nicolas Haderer, Romain Rouvoy, Christophe Ribeiro, and Lionel Seinturier. Apisense: Crowd-sensing made easy. *ERCIM News*, 93:28–29, 2013.

[21] Ramaswamy Hariharan and Kentaro Toyama. Project Lachesis: Parsing and Modeling Location Histories. In Max J Egenhofer, Christian Freksa, and Harvey J Miller, editors, *Geographic Information Science: Third International Conference, GIScience 2004, Adelphi, MD, USA, October 20-23, 2004. Proceedings*, pages 106–124. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[22] B Henne, C Kater, M Smith, and M Brenner. Selective cloaking: Need-to-know for location-based apps, 2013.

[23] Christian S Jensen, Hua Lu, and Man Lung Yiu. Location Privacy Techniques In Client Server Architectures. *Privacy in Location-Based Applications*, 5599:31–58, 2009.

[24] J K Laurila, Daniel Gatica-Perez, I Aad, Blom J., Olivier Bornet, Trinh-Minh-Tri Do, O Dousse, J Eberle, and M Miettinen. The Mobile Data Challenge: Big Data for Mobile Computing Research. In *Pervasive Computing*, 2012.

[25] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115, 2007.

[26] Xinxin Liu, Han Zhao, Miao Pan, Hao Yue, Xiaolin Li, and Yuguang Fang. Traffic-aware multiple mix zone placement for protecting location privacy. In *Proceedings of the IEEE INFOCOM 2012, Orlando, FL, USA, March 25-30, 2012*, pages 972–980, 2012.

[27] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 24, 2006.

[28] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. Ap-attack: A novel re-identification attack on mobility datasets. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MobiQuitous 2017, Melbourne, Australia, November 7 - November 10, 2017*, 2017.

[29] Open Street Map. Amenity information description: https://wiki.openstreetmap.org/wiki/key:amenity, 2018.

[30] Open Street Map. Download open street map dataset: https://wiki.openstreetmap.org/wiki/downloading_data, 2018.

[31] Kristopher Micinski, Philip Phelps, and Jeffrey S Foster. An Empirical Study of Location Truncation on Android. *Most'13*, 2013.

[32] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In *SenSys*, pages 323–336, 2008.

[33] Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Péter Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *MobiSys*, pages 55–68, 2009.

[34] Balaji Palanisamy and Ling Liu. Mobimix: Protecting location privacy with mix-zones over road networks. In *Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany*, pages 494–505, 2011.

[35] Michal Piorkowski, Natasa Sarafijanovic-djukic, and Matthias Grossglauser. CRAW- DAD data set epfl/mobility (v. 2009-02-24), 2009.

[36] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Differentially Private Location Privacy in Practice. *Most'14*, (October), 2014.

[37] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 1, pages 539–546. IEEE, 2015.

[38] Wahbeh H. Qardaji, Weining Yang, and Ninghui Li. Differentially private grids for geospatial data. In *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 757–768, 2013.

[39] Pierangela Samarati and Latanya Sweeney. Generalizing Data to Provide Anonymity when Disclosing Information. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '98, pages 188—-, New York, NY, USA, 1998. ACM.

[40] Pravin Shankar, Vinod Ganapathy, and Liviu Iftode. Privately querying location-based services with SybilQuery. *UbiComp*, page 31, 2009.

[41] Yu Zheng, Xing Xie, and Wei-Ying Ma. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data(base) Engineering Bulletin*, 2010.

[42] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering Personal Gazetteers: An Interactive Clustering Approach. In *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, GIS '04, pages 266–273, New York, NY, USA, 2004. ACM.