MOOD : MObility Data Privacy as Orphan Disease

Besma Khalfoun, Mohamed Maouche, Sonia Ben Mokhtar, Sara Bouchenak

▶ **To cite this version:**

Besma Khalfoun, Mohamed Maouche, Sonia Ben Mokhtar, Sara Bouchenak. MOOD : MObility Data Privacy as Orphan Disease. COMPAS, 2019, Biarritz, France. hal-03339301

HAL Id: hal-03339301
https://hal.science/hal-03339301v1

Submitted on 9 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ComPAS'2019 : Parallélisme/ Architecture / Système
LIUPPA - IUT de Bayonne, France, du 24 au 28 juin 2019

MooD : *MO*bility Data Privacy as *O*rphan *D*isease

Besma Khalfoun[1,2], Mohamed Maouche[1], Sonia Ben Mokhtar[1], and Sara Bouchenak[1]

[1]INSA Lyon – LIRIS, France {firstname.lastname}@insa-lyon.fr
[2]Ecole Nationale Supérieure d'Informatique Algiers, Algeria

Abstract
With the increasing development of hand held devices, Location Based Services (LBSs) became very popular facilitating users' daily life with a broad range of applications (e.g. traffic monitoring, geo-located search, geo-gaming). However, several studies have shown that the collected mobility data may reveal sensitive information about end users such as their work and home places, their gender, political, religious or sexual preferences. To overcome these threats, many Location Privacy Protection Mechanisms (LPPMs) were proposed in the literature. While the existing LPPMs try to protect most of the users in mobility datasets, there are always a subset of users who are not protected by any of the existing LPPMs. By analogy to medical research, there are orphan diseases, for which the medical community is still looking for a remedy. In this paper, we present MooD, a multi-LPPM user centric solution which main objective is to find a *treatment* to orphan users and protect them from re-identification attacks. Our experiments on the MDC dataset show that the percentage of protected users increased from 64 % to 98% on the considered dataset which is a promising result.

Keywords Geo-Privacy, Mobility Data, Location Privacy, Utility, Re-identification Attacks.

1. Introduction

Nowadays, the proliferation of mobile devices embedding GPS chips (e.g. smartphones, tablets, smart watches) has significantly contributed to the development of geolocatedservices, also named Location Based Services (LBSs). These services are useful for users' daily life as they allow them to localize nearby friends, discover their environment and request for places whenever they like and wherever they are. The downside is that huge amounts of information regarding users locations are being gathered and stored by third-party services. These services or other entities that may have access to the collected data (e.g., accidentally or through an attack) may exploit it fraudulently in order to infer and reveal sensitive information about individuals (e.g. home address, workplace, religious beliefs, sexual preferences, political orientations, social relationships). Consequently, it opens the door to several serious privacy threats. The most common threats are (1) re-identification attacks where anonymous mobility traces are re-associated to its originating user based on previously recorded data, (2) mobility prediction where users' next moves are anticipated, (3) extraction of user's places of interest (home, workplace, etc.) and (4) inference of social relationships (friends, coworkers, etc.). To tackle these threats, many Location Privacy Protection Mechanisms (LPPMs) were proposed in the literature. They transform mobility data relying on a wide list of techniques such as perturbation, generalization, and fake data generation [5].

To evaluate the effectiveness of LPPMs, a variety of privacy metrics are usually used and the resilience against re-identification attacks is one of them. Considering a protected mobility trace (i.e., a raw trace to which a given LPPM is applied), a re-identification attack tries to link the protected mobility trace to its owner based on past unprotected mobility data that the attacker has access to. There exist a variety of re-identification attacks in the state-of-the-art literature that differs in the way they model and analyse user mobility [4],[12]. However, when LPPMs are evaluated against re-identification attacks the focus is generally put on the protection of the crowd, i.e., protecting the larger proportion of users possible, and little attention is given to users that remain unprotected. Considering a set of state-of-the-art attacks and LPPMs at the disposal of a data security expert aiming at the protection of a given dataset, the question that the latter may ask is : What should be done with mobility traces that are subject to re-identification despite the use of LPPMs ?. These users are considered as *orphan* users and their protection needs to be addressed. In this paper, we present MooD (MObility Data Privacy as Orphan Disease), a user-centric approach to enforce location privacy using multiple LPPMs aiming at the protection of orphan users, i.e, users that are not protected against re-identification attacks while using any of the existing LPPMs. The originality of MooD is that it combines off-the-shelf LPPMs and applies for fine-grained protection. The LPPMs combination is realized with the application of various LPPMs on the same trace in the form of function composition, while the fine-grained protection implies the application of various LPPMs on contiguous sub-traces. MooD's mechanisms are driven by the resilience to state-of-the-art re-identification attacks and the data utility metrics set by the data security expert.
We evaluate MOOD by applying it to a real-life mobility dataset and comparing its performance to the application of single LPPMs. The results of our experiments show that MOOD is able to protect 98 % of the dataset whereas only 64% is protected by the considered single LPPMs.
The remaining of this paper is stuctured as follows. First, we present in Section 2, a background of re-identification attacks and location privacy protection mechanisms. Then, we describe our system design priciples in section 3. Further, in section 4, we proceed to an experimental evaluation of our solution. Finally, we conclude in section 5.

2. Background and Related Work

User Re-identification Attacks. A user re-identification attack aims at associating a protected mobility trace to its originating user based on users' past mobility. Two phases are necessary to run these attacks : a training phase and an attack phase. In the training phase, the attacker collects non-obfuscated mobility history of known users from several sources and builds users' mobility profiles. Several models have been used in the literature to characterize mobility profiles of users. The most common models include Points of Interests (POIs) (i.e., the set of meaningful places where users spent time) [12], Mobility Markov Chain (MMC) where states are POIs and edges reprensent the probability transition between states [4] or HeatMaps that aggregate user mobility over time across cells [9]. Then, in the attack phase, the attacker that receives an anonymous mobility trace, tries to re-associate it to the closest user profile among the learnt ones.
User Location Privacy Protection Mechanisms In order to mitigate location privacy threats, Location Privacy Protection Mechanisms (LPPMs) have been introduced in the literature. Formally, an LPPM is defined as a function. It takes as input a mobility trace T, which is 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, and produces as output a new obfuscated mobility trace T'. LPPMs differ in the way they alter the original

mobility data and in the guarantees they offer to end-users. These guarantees can be theoretical (e.g., k-anonymity [14] , differential privacy [3]) or practical (e.g., the resilience to known attacks).
Examples of such LPPMs include Geo-indistinguishability (Geo-I) [2] an instance of differencial privacy [3] in the context of location privacy. It is based on adding a random spatial noise (e.g Laplacian noise) to each GPS coordinate in the mobility trace. Trilateration based mechanism (TRL) [6] is a different way to generate dummies in online services. It is based on trilateration : when a user launches an LSS (i.e.Location Searching Service) query, looking for a restaurant or a gas station for example, the algorithm chooses randomly 3 assisted locations $l1$, $l2$ and $l3$ in a range of r from the real location l of the user. The service provider looks for nearby places according to the three reported locations and sends the result to the user. Finally, the user gets an accurate result by intersecting the result of the three sent locations using trilateration. Heatmap confusion (HMC) [10] is a combination of data perturbation technique and dummies generation. In HMC, the mobility trace of each user is represented as a heatmap. Then, the algorithm alters the given heatmap by making it look similar to the one of another user. The objective of such approach is to preserve a certain level of data utility and confuse an attacker that tries to re-identify users'mobility traces. Finally, HMC transforms back each obfuscated heatmap into a set of mobility traces. Moreover, recent works were proposed in the literature where user centric approach where was taken. Each individual is protected according to his mobility trace, its characteristics and his preferences in term of privacy. SmartMask is an example of that, [8]. It is a user centric framework, designed to automatically learn users privacy preferences under different contexts (e.g. location semantic, frequency of visits, duration of visiting a location, time period) and to provide a transparent and variable privacy control.

3. MooD Design Principles

3.1. Problem Illustration

Consider a data security expert that has to protect a given mobility dataset before its publication. The security expert has access to a set of LPPMs and to a set of user re-identification attacks found in the state-of-the-art literature. In order to assess the effectiveness of the LPPMs in front of the attacks, the expert may decide to run the re-identification attacks on the protected dataset and choose the LPPM that better protects her original dataset. In this direction, a discussion was undertaken in [9]. They selected three state-of-the-art LPPMs (Geo-I [2], W4M [1], Promesse [13]) on which three state-of-the-art re-identification attacks were launched (i.e., POI-Attack [12], PIT-Attack [4] and AP-Attack [9]). The results show that the proportion of users who are not protected with any single LPPM varies from 2% to 33% on four datasets. These users are considered as orphans and should be protected too.

3.2. Overview of MooD

MooD is a multi-LPPM user-centric approach. It can be used in a crowd-sensing campaign where users send their mobility traces periodically to a private proxy server or in data publishing use case where a data analyst aims at publishing users' traces.
The main objective of it is to protect the mobility trace of each user and in particular the minority of users who are not protected by any of the existing single LPPMs. The model architecture of MooD is depicted in Figure 1. It takes as inputs a mobility trace T , a set of LPPMs P and a set of User Re-identification attacks A. It returns an obfuscated trace Θ as an entire mobility trace T' or as multiple sub-traces $\{T'_1, T'_2, ..., T'_n\}$.

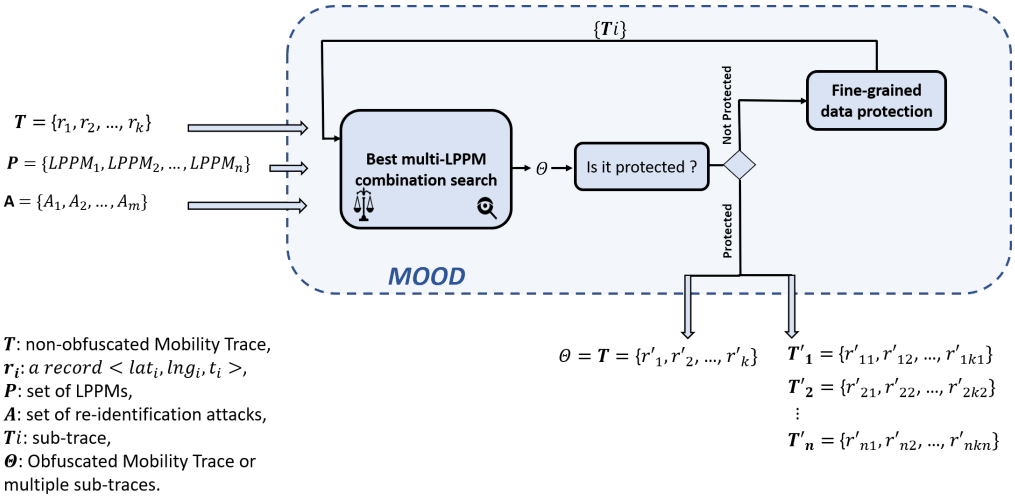

FIGURE 1 – MOOD Architecture

3.3. Best Multi-LPPM Combination Search

The Best Multi-LPPM Combination Search is the main component in our system. It takes as input a user's mobility trace T, a set of LPPMs P and a set of all considered re-identification attacks A. First, MOOD starts by applying all possible combinations of the considered LPPMs in an incremental and exhaustive manner so that the output mobility trace of the current LPPM becomes the input mobility trace of the following LPPM. The order of LPPMs is important since it is similar to function composition. For N = | P |, the number of possible combinations is $\sum_{i=1}^{N}(A_N^i)$, e.g. when N = 3, we get 15 possible combinations of LPPMs.
After that, once a mobility trace T is transformed by each composition of LPPMs separately, the re-identification attacks $\{ A_i \}_{i=1..m}$ are launched in order to evaluate the resilience of each combination of LPPMs and keep only ones that prevent from re-identification (if any). If all the re-identification attacks fail in re-association the obfuscated mobility trace Θ to its originating user, the privacy protection process is done and the user's mobility trace is protected by MOOD. In this case, a utility metric is computed ,i.e. *spatio-temporal* distortion metric [10]. basically, it measures the spatial distortion between non-obfuscated and obfuscated data and selects only the combination of LPPMs with the best utility (i.e. the lowest distortion). However, if at least one re-identification attack succeeds, it means that the user is still vulnerable. In this case, the mobility trace of the user is taken care by the next component.

3.4. Fine-Grained Data Protection

The fine-grained data protection is a complementary component in our system. It is launched when the user's mobility is protected by neither a single LPPM nor a combination of LPPMs. The idea we adopt is to split the original trace into a set of sub-traces and try to protect each sub-trace separately. Thus, we reduce the characterization of the user's mobility behavior because the longer the user's mobility trace the better its characterization. Several techniques of splitting traces can be used or combined. In our work, we opt for a fixed time slice where we split the trace after a fixed time duration (e.g every hour). The idea behind going towards fine-grained sub-traces derives from the fact that it is better to protect partly a user mobility data rather than not being able to protect it at all.

4. Experimental Evaluation

4.1. Experimental Setup

All the experiments were carried out in a computer running an Ubuntu 16.04 LTS OS with 5 GO of RAM and 3 cores of 1.8 GHz each. The different considered LPPMs and attacks are taken from a library *Accio* [11].

User Re-identification Attack Configuration. The three chosen re-identification attacks in this paper are : AP-attack, POI-attack, and PIT-attack. They are parameterized as follows : AP-attack has a cell size parameter fixed at 800 meters as a default value in [9]. POI-attack and PIT-attack require a diameter of the clustering area to extract Points of Interest (POIs), and a duration of time spent at a POI. They were respectively set to 200 meters and 1 hour as discussed in [9].

LPPM Configuration. To evaluate MOOD, we select three representative LPPMs : (1) Geo-I, (2) TRL and (3) HMC as single LPPMs. Each LPPM was configured as follows : Geo-I has ε as a privacy parameter which tunes the amount of noise added to the mobility data. (the lower ε the higher the protection). It was set to 0,01 which corresponds to a medium privacy level. TRL has a range of r from the real user's position where the fake locations are generated. The latter was set to 1 km. Finally, the cell size in heatmaps of HMC was set to 800 meters which concords with the value used in [10].

Utility Metric We used the spatio-temporal distortion metric (STD) to evaluate the effectiveness of both single and multiple LPPMs combinations [10]. As defined in Equation 1, the $\mathcal{STD}$ is the average distance between each record of the obfuscated trace T' and its temporal projection into T. Specifically, we search for $r_i = (lat_i, lon_i, t_i)$ and $r_{i+1} = (lat_{i+1}, lon_{i+1}, t_{i+1})$ in T such as $t_i \leq t_x \leq t_{i+1}$, then compute r_e the interpolation with the ratio $(t_x - t_i)/(t_{i+1} - t_i)$.

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

4.2. Mobility Dataset Description

In our experiments, we used MDC dataset [7] that contains the mobility of 141 users in the city of Geneva (Switzerland). In the first and the second experiments, the 30 most active successive days were considered. The mobility trace of each user was chronologically split into 15 days period as a training set and the remaining 15 days period as a testing set. For the fine-grained data protection, precisely in the testing set, each trace was split into sub-traces of a one day period to simulate the scenario of a crowdsensing application where users send their data daily.

4.3. Experimental Results

Resilience to Single Re-identification Attack. As a first experiment, we want to showcase the problem of orphan users when a single attack is used by the data analyst. We consider a set of state-of-the-art LPPMs (n = 3) and we select AP-attack as - the most powerful attack - in the literature. The result, as depicted in Figure 2, shows that 96 out of 141 users were not protected naturally (i.e. without applying any LPPM), 10 out of 141 users who were not protected by a single LPPM became protected with MOOD. Therefore, at the end of the experiment, our system was able to protect all the dataset.

Resilience to Multiple Re-identification Attacks. As *User Re-identification attacks* do not necessarily re-identify the same users, we decide to consider a stronger attacker with multiple attacks (m=3) (POI-attack, PIT-attack, and AP-attack). As shown in Figure 3, 107 out of 141 users are not protected naturally when no LPPM is applied. 51 out of 141 users are not protected

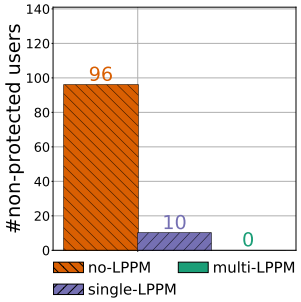

FIGURE 2 – Resilience to one attack – Single LPPM vs. multi-LPPMs

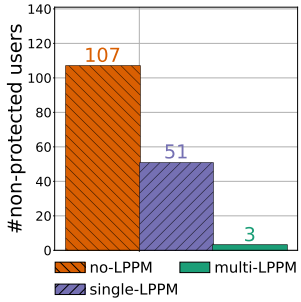

FIGURE 3 – Resilience to multiple attacks – Single LPPM vs. multi-LPPMs

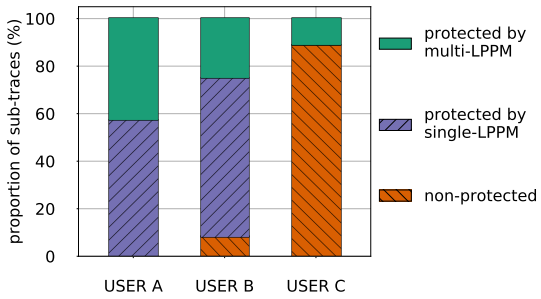

FIGURE 4 – Resilience to multiple attacks – Fine-grained protection

when a single LPPM is applied whereas only 3 among the remaining 51 users are re-identified with MOOD. It means that we are able to increase the percentage of protected users from 64 % to 98 % over the whole dataset. Furthermore, we evaluated the impact of MOOD on data utility. We consider three acceptable limits of the spatial-temporal distortion : low, medium, high. The table 1 shows that 27%, 94% and 98% of protected users with low, medium or high spatial distortion respectively. Depending on the degree of distorted data, we can imagine several scenarios of data publishing and crowd-sensing applications using MOOD. For instance, measuring the level of noise or pollution in a specific city or for weather forecasting when the distortion is high and the accuracy of the protected data is not important.

Spatial-temporel distortion limit (m)	500 (Low)	1000 (Medium)	5000 (High)
Pourcentage of protected traces with MOOD (%)	27	94	98

TABLE 1 – Percentage of protected mobility traces with different levels of the spatio-temporal distortion metric.

Fine-Grained Data Protection. In this experiment, we zoom on the three left unprotected users by MOOD, {A, B, C}. We split their mobility traces into sub-traces of 24 h. Overall the users, we obtain 28 sub-traces. Figure 4 shows that there are 9 unprotected sub-traces, 12 protected with single LPPM and 7 protected with Multi-LPPM. Thus, we can see that user A became totally protected. User B almost protected, whereas User C is still non-protected. Thus, the granularity of considered traces has an impact on privacy protection.

5. Conclusion

In this paper, we presented MOOD a user centric system based on composition of multiple LPPMs. Its main objective is to protect the minority of orphan users who are not protected by any single LPPM. The experiments show that MOOD is resilient to re-identification attacks and can achieve a high level of privacy protection. Furthermore, due to the lack of space, we would like to test MOOD on other datasets. In addition, our work is extensible and other LPPMs and attacks can easily be integrated. Another futur open direction is to test more sophisticated techniques in the fine-grained data protection. For instance, a mobility trace can be split based on the number of POIs or according to a fixed time gap between locations.

References

1. Abul (O.), Bonchi (F.) et Nanni (M.). – Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, vol. 35, n8, 2010, pp. 884–910.
2. Andrés (M. E.), Bordenabe (N. E.), Chatzikokolakis (K.) et Palamidessi (C.). – Geo-indistinguishability : differential privacy for location-based systems. – In *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*, pp. 901–914, 2013.
3. Dwork (C.). – Differential privacy. *Encyclopedia of Cryptography and Security*, 2011, pp. 338–340.
4. Gambs (S.), Killijian (M.-O.) et del Prado Cortez (M. N.). – De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, vol. 80, n8, 2014, pp. 1597–1614.
5. Gupta (R.) et Rao (U. P.). – An exploration to location based service and its privacy preserving techniques : A survey. *Wireless Personal Communications*, vol. 96, n2, 2017, pp. 1973–2007.
6. Huang (Y.), Cai (Z.) et Bourgeois (A. G.). – Search locations safely and accurately : A location privacy protection algorithm with accurate service. *J. Network and Computer Applications*, vol. 103, 2018, pp. 146–156.
7. Laurila (J. K.), Gatica-Perez (D.), Aad (I.), J. (B.), Bornet (O.), Do (T.-M.-T.), Dousse (O.), Eberle (J.) et Miettinen (M.). – The Mobile Data Challenge : Big Data for Mobile Computing Research. – In *Pervasive Computing*, 2012.
8. Li (H.), Zhu (H.), Du (S.), Liang (X.) et Shen (X. S.). – Privacy leakage of location sharing in mobile social networks : Attacks and defense. *IEEE Transactions on Dependable and Secure Computing*, vol. 15, n4, 2018, pp. 646–660.
9. Maouche (M.), Mokhtar (S. B.) et Bouchenak (S.). – Ap-attack : A novel user re-identification attack on mobility datasets. – In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems : Computing, Networking and Services, Melbourne, Australia, November 7-10, 2017.*, pp. 48–57, 2017.
10. Maouche (M.), Mokhtar (S. B.) et Bouchenak (S.). – Hmc. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, n3, 2018, pp. 1–25.
11. Primault (V.), Maouche (M.), Boutet (A.), Mokhtar (S. B.), Bouchenak (S.) et Brunie (L.). – ACCIO : how to make location privacy experimentation open and easy. – In *38th IEEE International Conference on Distributed Computing Systems, ICDCS 2018, Vienna, Austria, July 2-6, 2018*, pp. 896–906, 2018.
12. Primault (V.), Mokhtar (S. B.), Lauradoux (C.) et Brunie (L.). – Differentially private location privacy in practice. *arXiv preprint arXiv :1410.7744*, 2014.
13. Primault (V.), Mokhtar (S. B.), Lauradoux (C.) et Brunie (L.). – Time distortion anonymization for the publication of mobility data with high utility. *CoRR*, vol. abs/1507.00443, 2015.
14. Sweeney (L.). – k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, n05, 2002, pp. 557–570.