Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets

Antoine Boutet, Sonia Ben Mokhtar, Vincent Primault

Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets

Antoine Boutet, Sonia Ben Mokhtar, Vincent Primault
University of Lyon, LIRIS, CNRS,
INSA-Lyon, UMR5205, F-69621, France
{antoine.boutet,sonia.benmokhtar,vincent.primault}@liris.cnrs.fr

Abstract

The widespread adoption of handheld devices (e.g., smartphones, tablets) makes mobility traces of users broadly available to third party services. These traces are collected by means of various sensors embedded in the users' devices, including GPS, WiFi and GSM. We study in this paper the mobility of 300 users over a period up to 31 months from the perspective of the above three types of data and with a focus on two cities, i.e., Lausanne (Switzerland) and Lyon (France). We found that users' mobility traces, no matter if they are collected using GPS, WiFi or GSM antennas, are highly unique. We show that on average only four spatio-temporal points from the WiFi, GSM and GPS traces are enough to uniquely identify 94% of the individuals, on both datasets. In addition, we show that using the temporal dimension (i.e., whether users move or are in a meaningful location such as their home or their working place) drastically improves the capacity to uniquely identify them compared to when only exploiting the spatial dimension (by 14% on average). In some cases, using the temporal dimension alone can represent a better mobility footprint than the spatial dimension to discriminate users. We further conduct a de-anonymisation attack to assess how mobility traces can be re-identified, and show that almost all users can be de-anonymised with a high success rate. Finally, we apply different location privacy protection mechanisms (LPPMs), including spatial filtering, temporal cloaking, adding spatial noise to mobility data, or using generalisation, and analyse the impact of these mechanisms on both the uniqueness of users' mobility traces and the outcome of the de-anonymisation attack. We show that spatially obfuscating mobility data is not enough to protect users, and that classical LPPMs are not able to protect users against a de-anonymisation attack. We finally conclude this paper by drawing some insights towards future spatio-temporal LPPMs.

1 Introduction

The large adoption of mobile devices with embedded geolocation capabilities makes it possible to track the movements of a large number of individuals during their daily life. These mobility traces have a huge commercial value [1] and consequently raise increasing interest from the one hand and open the door to increasing threats on the other hand. These traces can be collected using various sensors embedded in users' handheld devices. While the Global Positioning System (GPS) is largely exploited to identify the location of users due to its high precision, the WiFi and the GSM can also be leveraged to track individuals' mobility [2]. Indeed, MAC addresses of WiFi access points or identifiers of GSM antennas users are associated to can be easily mapped to GPS coordinates using public repositories (e.g., WiGLE [1] or Google [2]).

Following the seminal work by De Montjoye and al. [3] that analysed user mobility traces inferred from call logs,

[1] WiGLE: Wireless Network Mapping, `http://wigle.net`

[2] Google Maps Geolocation API: `https://developers.google.com/maps/`

we present in this paper a similar study performed on user mobility traces coming from three other types of data, namely GPS, WiFi and GSM traces. Specifically, we used two datasets comprising 200 and 100 users spanning over 31 and 15 months, respectively. Our results show that on average only four random spatial and spatio-temporal points from the WiFi, the GSM and the GPS data collections are enough to uniquely identify 94% of the individuals with the two datasets. Moreover, we show that considering temporal-only mobility traces (i.e, whether users are moving or inside meaningful locations such as their home or work place) can represent a mobility footprint able to highly discriminate between individuals (i.e., uniquely identify up to 84% of the individuals). In addition, we show that the temporal information improves the capacity to uniquely identify individuals by 14% on average compared to only considering spatial information. Furthermore, by analysing the degree of uniqueness of individual users, we show that the latter is heterogeneous (i.e., some users have more discriminative mobility patterns than others). We also compared the uniqueness given by two different models and we show that the probabilistic uniqueness assessment proposed by De Montjoye and al. [3] gives an upper bound of the uniqueness compared to the deterministic assessment proposed by Zang and Bolot [4]. Finally, we analyse the impact of applying classical location privacy protection mechanisms (LPPMs), namely spatial filtering, temporal cloaking, spatial noise addition providing ϵ-differential privacy properties, and generalization ensuring k-anonymity, on the uniqueness of mobility traces. We show that obfuscating only the spatial dimension of mobility data is not enough to reduce the uniqueness of users.

Measuring the uniqueness does not mean re-identification of users. Indeed, pseudo-anonymised mobility traces themselves do not disclose the identity of a user. However, using external knowledge can lead to infer the identity of users [5, 6, 7]. Instead of analysing the re-identification of users which requires external knowledge, we analysis here how users can be de-anonymised from their mobility traces. To achieve that, we also conduct a de-anonymisation attack [8] trying to associate each individual inside a training set of mobility traces to its anonymous counterpart inside a testing set. We show using our datasets that geolocated data can be almost fully de-anonymised. Furthermore, we show that applying LPPMs based on noise or generalization on mobility traces fail to protect users against this de-anonymisation attack.

In this paper, we seek to answering several questions such as: is the uniqueness of mobility traces from GSM, GPS, and WiFi similar to the one observed on previous studies? ; are the temporal and the spatial dimensions similarly discriminating? ; does the uniquess vary from one user to another? ; do all uniqueness assessment models provide similar results? ; what is the impact of LPPMs on the uniqueness of individuals? ; do LPPMs efficiently protect users against a de-anonymisation attack? We hope that the observations we make while answering these questions can lead to the development of more effective LPPMs in the future. To summarise, the takeaways of this study are:

- User mobility traces extracted from GPS, WiFi and GSM data are highly unique, which generalises the results of the study performed in [3] on call logs.
- Temporal data is as discriminative as spatial data in human mobility traces, which shall be considered in the development of future LPPMs.
- The uniqueness degree of mobility traces is user dependent while most existing LPPMs are statically configured for all users. Consequently personalisation shall thus be introduced in future LPPMs.
- Applying mechanisms to reduce the uniqueness of mobility traces drastically impacts the utility of the data, which comforts previous study [4, 2].
- Probabilistic uniqueness assessment based on the methodology proposed De Montjoye and al. [3] gives an upper bound of the uniqueness compared to a deterministic approach proposed by Zang and Bolot [4].
- Depending on the nature of the LPPMs, obfuscating the data collection, the results in term of uniqueness can be very different.
- Applying a de-anonymisation attack leads to re-identify users with a high success rate, even if the raw data is obfuscated by classical LPPMs.

The remaining of this paper is organised as follow. Section 2 presents background and related works. Section 3 and 4 then describe the methodology and the evaluation, respectively. Finally, Section 5 discusses and concludes this paper.

2 Background

The idea of finding criteria for uniquely identifying users is not new. In criminal investigations, finding seven points of minutiae in a fingerprint is commonly used to uniquely identify an individual and provide matching evidence. In computer science, the research community has investigated various types of user traces that may act as digital fingerprints such as the personalised configurations of mobile devices [9] or Web browsers [10], the logs of in-car sensors [11], or the writing style of users on the Web [12]. Recently, the uniqueness of human mobility traces has been extensively analysed by De Montjoye and al. [3]. In this paper, the authors analysed mobility traces coming from the call logs of 1.5M users at the scale of a country. Their results show that only four random spatio-temporal points were enough to uniquely identify 95% of the individuals of the dataset. In this paper we run a complementary study by using mobility traces coming from three different data sources, i.e., GPS, WiFi and GSM logs.

Zang and Bolot [4], also show that mobility patterns of users often make them unique within a large population but with another model to assess the uniqueness. Instead of using a set of random mobility patterns for each user to quantify the uniqueness (i.e. probabilistic approach followed by De Montjoye and al. [3]), these authors used a deterministic approach which evaluates the uniqueness of the mobility pattern composed of the most frequently visited locations by the associated user. In this paper, we compared both models and show that the probabilistic one gives an upper bound of the uniqueness compared to the deterministic one.

Finally, previous studies [4, 3, 13, 2] show that reducing the uniqueness requires severe reductions of the spatio-temporal granularity which limits the usability of the data. In this paper, we confirm this observation with data from other sensors.

Uniqueness does not mean re-identification, since pseudonymity is used to avoid revealing the real identity of users (i.e., only mobility traces of users do not disclose their identity). However pseudonymity alone is not enough to guarantee anonymity [14]. Furthermore, inferring user identity may become possible by leveraging external knowledge. Using crowdsourcing or a cross-database methodology, recent works have demonstrated the re-identification risks from smartphone metadata [5], social network data [6], or movie databases [7]. To address the challenge of location privacy, many LPPMs have been recently proposed in the literature [15, 16, 17, 18]. These LPPMs apply different schemes to obfuscate the location information of users. The two most adopted privacy guarantees provided by LPPMs are k-anonymity [19] and ϵ-differential privacy [20]. While the former hides each user within a cloaking area containing at least $k-1$ other users, the latter disturbs mobility traces in such a way that it theoretically bounds the impact of the presence or absence of a single element of the dataset. For instance, [21] describes a protection mechanism providing k-anonymity by relying on a centralised anonymisation proxy. In this protocol, the proxy receives all queries from the user, generates a cloaking area for each of them before sending the obfuscated query to a location-based service (LBS). Then the proxy extracts a more precise answer from the response coming from the LBS, before returning the results to the user. In [22], the authors removed the dependency to a trusted proxy by presenting a fully distributed protection mechanism dynamically building cloaking areas of at least k users during their mobility. To ensure k-anonymity, many approaches have been proposed to build cloaking areas including generalization and suppression [23], condensation [24] or space translation [25]. For instance, Wait 4 Me [25] (W4M for short) ensures k-anonymity by transforming the GPS coordinates of a moving object to a cylindrical volume representing the trajectory of this object, where the radius δ of the cylinder represents the possible location imprecision (i.e., we do not know exactly where the object is located within the cylinder). In this context, k-anonymity is guaranteed if k objects move within the same cylinder.

Geo-Indistinguishability (GEO-I) [16] was proposed as a generalisation of differential privacy applied to mobility data. The guarantee can be enforced by adding calibrated noise drawn from a two-dimensional Laplace distribution. GEO-I has been successfully applied in an online context

when a user is querying an LBS in real-time, and in an offline context when an entire dataset gathering the mobility traces of a set of users has to be protected before to be released. The practical impact of GEO-I has been studied in [26]. Authors analyse the effect of this LPPM on various location privacy attacks. They show that by adapting the algorithms to the underlying LPPM, it is possible to decrease or counteract most of its effects.

While several LPPMs have been proposed in the literature, their impact on the uniqueness of mobility traces have not been evaluated. In addition, a better understanding of the characteristics of mobility fingerprints can lead to improve LPPMs. For instance, most of the above LPPMs focus on obfuscating the spatial dimension while neglecting the temporal one [16]. Furthermore, most of them are static and do not evolve according to the considered user. In this paper, we highlight these limitations in numbers.

To infer the identity of a particular individual behind a set of mobility traces, Gambs and all [8] proposed a de-anonymisation attack. This attack is based on two phases, the first one is used to build a Mobility Markov Chain model on a training set which is used in the second one to re-identify users in a testing set. Freudiger and all [27] also try to re-identify users of geolocated datasets. In this study, the pair of POIs home/work is inferred and used as pseudo-identifier to de-anonymise users. As far as we know, no previous works analyse the impact of GEO-I and W4M on de-anonymisation attack. In this paper, we implement such an attack and analyse the impact of classical LPPMs on its outcome.

3 Methodology

This section starts with the presentation of the considered datasets and the methodology to extract mobility traces from these data collections. We then describe how the uniqueness is quantified. Finally we present the considered de-anonymisation attack and the LPPMs used to protect the data.

3.1 Datasets

This work was performed using the PRIVAMOV and the Mobile Data Challenge (MDC) mobile phone datasets. Both datasets contain data about users during their daily life, captured through different modalities (i.e., communication, location, motion, application usage, etc.). Table 1 displays information on these datasets and the mobility-based data collections considered in this work. More precisely, the considered data collections gather information from the GSM, the WiFi, and the GPS sensors.

Dataset	Users	Period	Sensor	#Records
PRIVAMOV	100	10/2014 - 01/2016	WiFi	25,655,480
			GSM	8,076,512
			GPS	156,041,576
MDC	200	10/2009 - 03/2011	WiFi	53,432,599
			GSM	50,895,615
			GPS	11,077,061

Table 1: Our uniqueness study uses the PRIVAMOV and the Mobile Data Challenge (MDC) mobile phone datasets with data collections captured through different sensors.

The PRIVAMOV dataset involves 100 students and staff from various campuses in the city of Lyon [28] equipped with smartphones running a data collection software. The data collection took place from October 2014 to January 2016. The MDC dataset involves around 200 volunteers [29, 30]. The data collection took place from October 2009 to March 2011 in the Lake Geneva region, Switzerland.

No filtering scheme was applied on the raw data contained in the PRIVAMOV dataset. However, a privacy protection scheme based on k-anonymity has been performed on the raw data before releasing the MDC dataset. This concerns all the location data including the information from the GPS, the WiFi, and the GSM sensors. Unfortunately, as described in [29], this privacy preserving operation includes many manual operations which are difficult to fully understand and reproduce. Yet, as shown later in Section 3.2, the impact of these manipulations are noticeable through different data distributions.

3.2 Mobility trace extraction

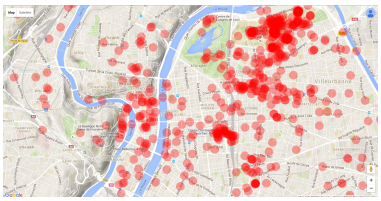
From the GSM, WiFi and GPS data collections, we extracted and built mobility traces. Each mobility trace is a list of spatio-temporal points belonging to a given user. The temporal component is the time of an interaction (the

Dataset	Identifier	Global	Sub-area
PRIVAMOV	GSM Antenna	19,033	3,816
	WiFi Access Point	407,690	168,741
	Point Of Interest	756	204
MDC	GSM Antenna	100,168	589
	WiFi Access Point	566,390	6,018
	Point Of Interest	2,816	343

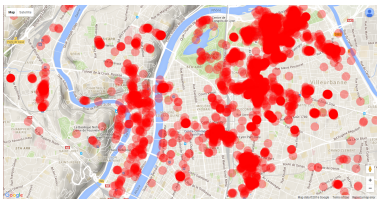
Table 2: Human mobility traces are built using GSM antennas and WiFi access points users encounter, and extracted POI for the whole datasets and a restricted sub-area.

temporal resolution is by default of one hour) and the spatial component depends on the data collection. For instance, for mobility traces coming from the GSM data collection (noted T_{gsm}), the spatial component of each point corresponds to the identifier of the GSM antenna to which the user is connected. Further, for mobility traces coming from the WiFi data collection (noted T_{wifi}) the spatial component corresponds to the MAC address of the WiFi access point. Finally, for mobility traces coming from the GPS data collection (noted T_{gps}), we extracted Points of Interest (POIs) from each trace, assigned an identifier to each unique POI inside the whole dataset and used this identifier as the spatial component. This means that for GPS traces, we only have points if the user is inside a POI.

To compute points of interest, we used a methodology similar to [31]. The idea behind this method is to identify restricted areas where users stay more than a specific duration. More precisely, POIs can be extracted using a simple spatio-temporal clustering algorithm parametrised with a maximum POI diameter d and a minimum stay time t. This POIs extraction is done in two clustering steps, the first one identifies POIs for each user and the second one assigns identifiers to unique POIs, thus allowing to identify POIs shared by several users. In this paper, by default, we use a diameter of 250 meters ($d = 250$) and a stay time of 30 minutes ($t = 30$). For instance, Figure 1 illustrates the POIs of users for the PRIVAMOV dataset in the Lyon sub-area for $t = 30$ minutes (a) and $t = 20$ minutes (b). Obviously, a shorter temporal resolution gives more POIs. Darker spots represent areas where more users generate and share common traces, and consequently less unique traces. We assessed the impact of both spatial and temporal resolutions in the POI extraction (parameter d and t) on the uniquenesses in Section 4.3.



(a) d=30 minutes



(b) d=20 minutes

Figure 1: Points of interest (POI) identify specific locations where a user stays in the same place longer than a specific delay d. These figures show the POIs of users across the Lyon area according to a delay considered to identify POIs of 30 and 20 minutes for figures (a) and (b), respectively. A shorter d provides more POIs.

Mobility traces do not contain duplicate entries at a given time resolution. For instance, if a user stays at the same location during one hour, even if the data collection software collects multiple records stating that its smartphone is associated to the same WiFi access point during this time slot, only one entry will be present in her mobility traces T_{wifi} at this given time. Conversely, if a user moves and gets connected to different WiFi networks, the resulting mobility trace T_{wifi} will contain several entries (i.e., one per WiFi access point associated to her smartphone). As a consequence, the number of points at a given hour reflects if the user is static (i.e., one point) or mobile (i.e., multiple points).

To evaluate separately the impact of the spatial and the temporal information, we also build spatial only and temporal only mobility traces. The spatial mobility traces for one user contains only the set of GSM antenna identifiers, WiFi access point's MAC addresses, or POI identifiers for traces from the GSM, WiFi and GPS data collections, respectively. The temporal mobility traces, in turn, are built to reflect when a user is moving or static. To achieve that, the temporal mobility trace of a user contains a set of

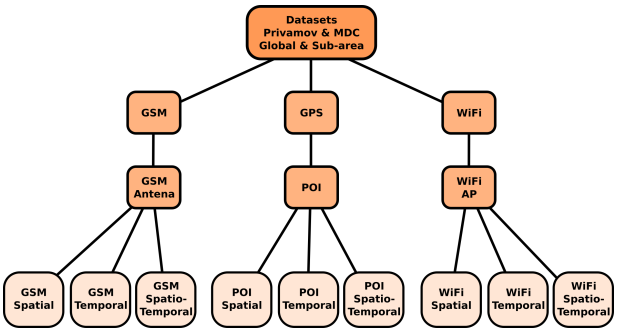


Figure 2: Methodology to build the mobility traces from the GSM, GPS, and WiFi data collections.

times (i.e., hours) where this user has met at least 3 WiFi access points or 3 GSM antennas for WiFi and GSM data collections, respectively. For POI-based temporal mobility traces, the set contains the times when a user is inside a POI.

Lastly, we also built mobility traces gathering only spatio-temporal points contained in a restricted sub-area. We consider the Lyon area for the PRIVAMOV dataset and Lausanne area for the MDC dataset excluding suburbs in both cases. Table 2 shows the number of unique GSM antennas, WiFi access points and POIs found inside each dataset, while Figure 2 depicts the methodology to build the mobility traces from the raw data collections.

To deeper analyse the resulting mobility traces, Figure 3 shows various Complementary Cumulative Distribution Function (CCDF, defined as $P(X > x)$) for both datasets. Figures 3a-3c-3e-3g show the number of unique GSM antennas, WiFi access points and POIs per user. Figures 3b-3d-3f-3h depict the number of unique users identified per GSM antenna, WiFi access point and POI. These tail distributions show that most of GSM antennas, WiFi access points and POIs have been visited only by one user. For instance, 75% of the PRIVAMOV WiFi access points have been seen by only one user. Conversely, mobility traces of most of the users are composed of several GSM antennas, WiFi access points and POIs. Interestingly, fewer MDC users have a small number of POIs, GSM antennas and WiFi access points than PRIVAMOV users. This difference is certainly due to the privacy protection scheme applied on the MDC dataset compared to the PRIVAMOV one. Finally, these figures show that users have been attached to more WiFi access points than GSM antennas, and the number of POIs is lower than the two others.

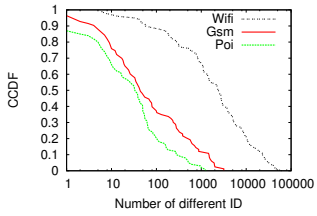
3.3 Uniqueness assessment

To quantify the uniqueness, we use the same methodology as presented in [3]. More precisely, as previously described, datasets contain for each user three mobility traces $T_{gps}, T_{wifi}, T_{gsm}$. These traces list the spatio-temporal points containing respectively the identifiers of POIs that the user has visited, MAC addresses of WiFi access points and identifiers of GSM antennas that the user was connected to, associated with the time of the interaction. For each mobility trace T, we evaluate the uniqueness of a given sub-trace I_p of p randomly chosen spatio-temporal points. A sub-trace I_p is said to be unique if only one user has $I_p \in T$. To measure this uniqueness, we performed a brute force search of users who have the p points composing I_p in their mobility trace T. The size of this set of users sharing the same I_p, noted k, characterizes the uniqueness of the sub-trace I_p. If $k = 1$, the sub-trace is unique. The uniqueness of traces is estimated as the percentage of 2500 random sub-traces that are unique given the p points composing them. We use the same methodology to evaluate the uniqueness of spatial or temporal only mobility traces. In this case, the sub-trace I_p contains spatial or temporal points, respectively. As reported in Section 3.2, we consider both no restriction for the chosen p points and specific sub-areas focusing on denser urban areas. Additionally, we also consider the case where the sub-trace I_p contains multi-sensor information and gathers spatial points from the GPS, WiFi, and the GSM data collections at once.

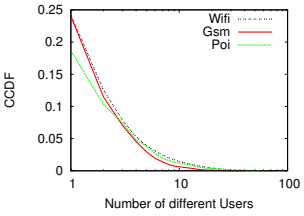
Finally, we also consider the deterministic model proposed in [4] to quantify uniqueness. In this model, the sub-trace I_p of a user is composed of the most frequent points (i.e., POIs, WiFi access points, and GSM antennas) that the user has visited the most frequently.

3.4 De-anonymisation attack

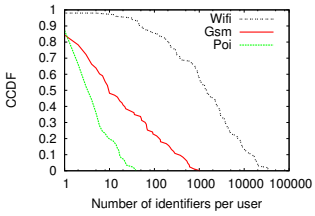
We also implement and conduct a de-annonymization attack similar to the one proposed by Gambs and all [8]. This attack aims at inferring the identity of a particular individual behind a set of mobility traces. More precisely,



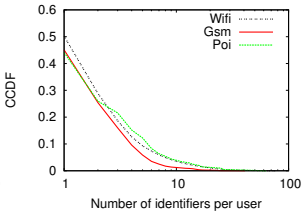
(a) PRIVAMOV - # of Identifiers



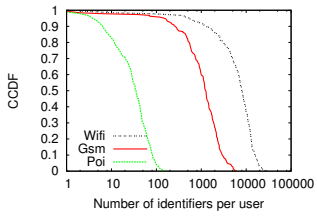
(b) PRIVAMOV - # of Users



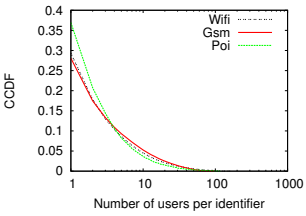
(c) PRIVAMOV (Lyon) - # of Identifiers



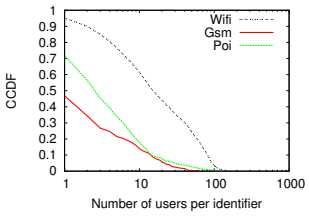
(d) PRIVAMOV (Lyon) - # of Users



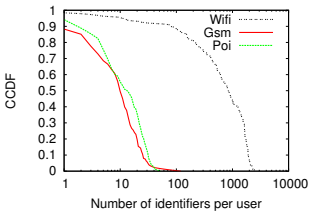
(e) MDC - # of Identifiers



(f) MDC - # of Users



(g) MDC (Lausanne) - # of Identifiers



(h) MDC (Lausanne) - # of Users

Figure 3: The tail distributions (i.e., Complementary Cumulative Distribution Function defined as $P(X > x)$) of the number of different GSM antennas, WiFi access points, and POI per user (Figures 3a-3c-3e-3g) ; and the number of unique users per GSM antenna, WiFi access point, and POI (Figures 3b-3d-3f-3h) for both a restricted sub-area (i.e., Lyon for Privamov and Lausanne for MDC) and for the whole dataset.

the mobility traces are split into a training and a testing set, and the attack tries to map each user of the testing set to the corresponding user in the training set. In our experiment, the training set of each user contains 70% of their first spatial points, and the testing set contains the remaining 30%. Only a limited knowledge is preserved from the training set. Specifically, we build for each user u the set Top_n which contains the Top n most frequently visited POIs, WiFi access points, and GSM antennas from the training set. Then, to re-identify a user, we measure the cosine similarity distance between all Top_n sets and the considered testing set. The re-identification is considered successful if only one user has the greatest similarity distance, otherwise the considered testing set is not associated to a particular user as the confidence of the mapping is not entirely undoubted. The outcome of the attack is measured through the precision and the recall metrics according to the value of n. Let us define U the set of users, $M(n)$ the set of users re-identified through the de-annonymization attack, and $C(n)$ the set of users correctly re-identified. Precision and recall are then defined as follow:

$$Precision(n) = \frac{|C(n)|}{|M(n)|}, \qquad Recall(n) = \frac{|C(n)|}{|U|}$$

The precision quantifies the accuracy of the re-identification while the recall assesses its completeness.

3.5 Location Privacy Protection Mechanisms (LPPM)

Lastly, we also quantify the impact of different LPPMs on the uniqueness and the outcome of the de-anonymisation attack on the resulting obfuscated mobility traces of users. We considered four different LPPMs, a naive spatial filter-based mechanism, a temporal cloaking, an obfuscation mechanism providing ϵ-differential privacy properties (i.e., Geo-Indistinguishability), and an obfuscation mechanism enforcing k-anonymity (i.e., Wait 4 Me).

3.5.1 Spatial Filtering

We consider here a mechanism that filters out the identifiers of POIs, GSM antennas or WiFi access points that have been only visited by f different users or less.

3.5.2 Temporal Cloaking

In this mechanism, we replace timestamps by a coarser grain temporal information. To achieve that, we vary the temporal resolution of the mobility traces from 1 hour up to 1 month.

3.5.3 Geo-Indistinguishability

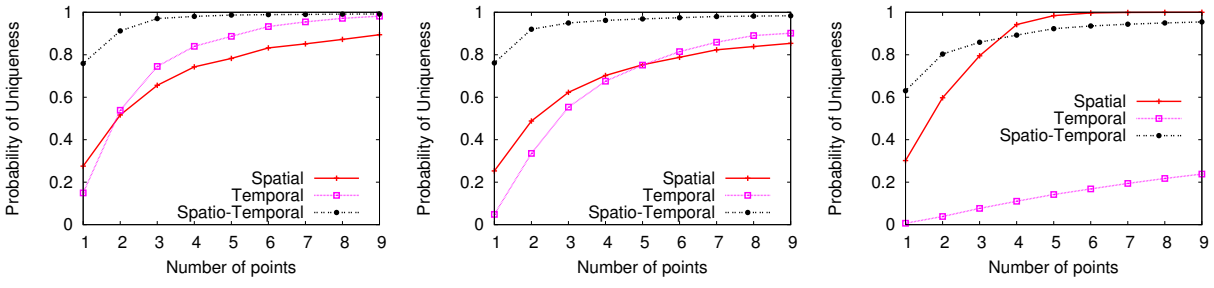
GEO-I [16] ensures ϵ-differential privacy by adding a calibrated noise drawn from a two-dimensional Laplace distribution. This LPPM takes a parameter names ϵ (expressed in meters^{-1}) determining the amount of noise to add (the smaller the value of ϵ, the higher the amount of noise added to the raw data). As reported in [32], we consider three values of ϵ: 0.01, 0.004, and 0.001, defining a low, medium, and high noise injection, respectively. As this LPPM obfuscates GPS coordinates, we apply this mechanism only on POI-based mobility traces.

3.5.4 Wait 4 Me

W4M [33] ensures k-anonymity by transforming the GPS coordinates of a moving object to a cylindrical volume of radius δ representing the possible location imprecision of this object. To achieve that, W4M can temporally and spatially move input records of raw mobility traces, as well as insert and delete records. We configure W4M with the following parameters: δ = 200 meters (i.e., the uncertainty), k = 2 (i.e., the anonymity level), $Max_{Trash} = 10\%$ of the dataset's size (i.e., the global maximum trash size), and $max_{radius} = 1000$ meters (i.e., the initial maximum radius used in clustering). This means that at any time, any two traces in the protected dataset are in a cylinder that has a 200 meters diameter. To comply with the usage restriction [33] (i.e., only 10,000 GPS records per user), we prepared a specific PRIVAMOV dataset containing only one record every minute and where records are randomly removed for users that exceeds this limit (i.e. 22 users over 92). As W4M obfuscates GPS coordinates, we only applied this mechanism on POI-based mobility traces.

4 Experimental Evaluation

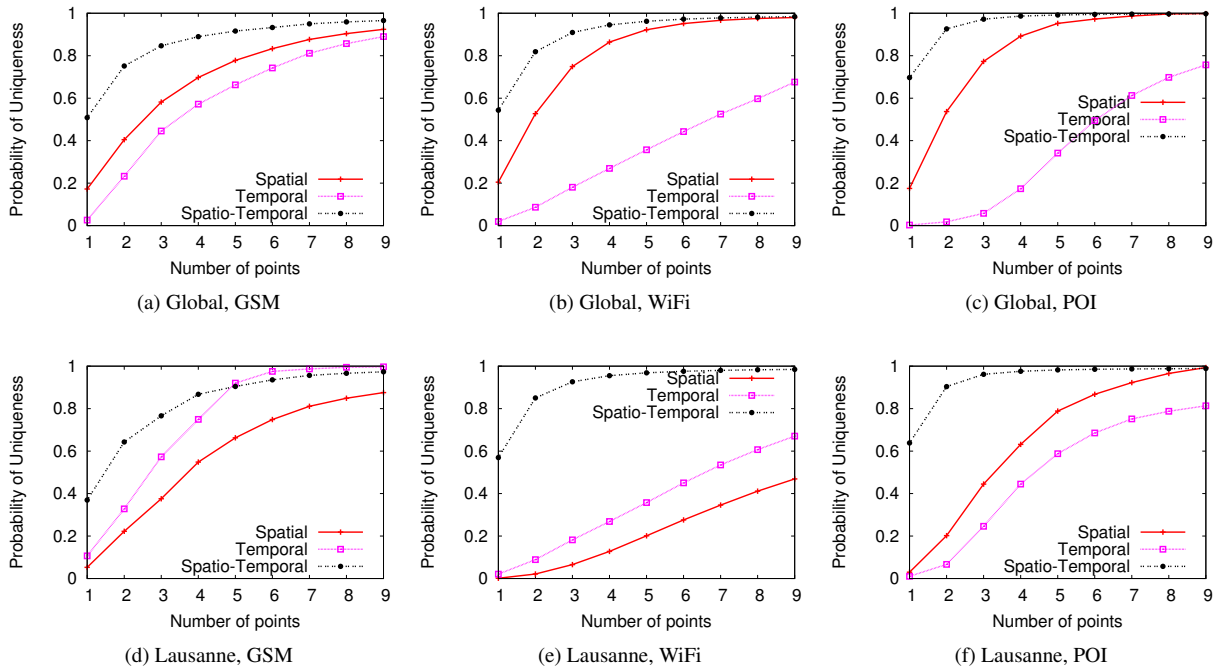
This section exhaustively evaluates the uniqueness of mobility traces before analysing the impact of different loca-



(a) Global, GSM (b) Global, WiFi (c) Global, POI

(d) Lyon, GSM (e) Lyon, WiFi (f) Lyon, POI

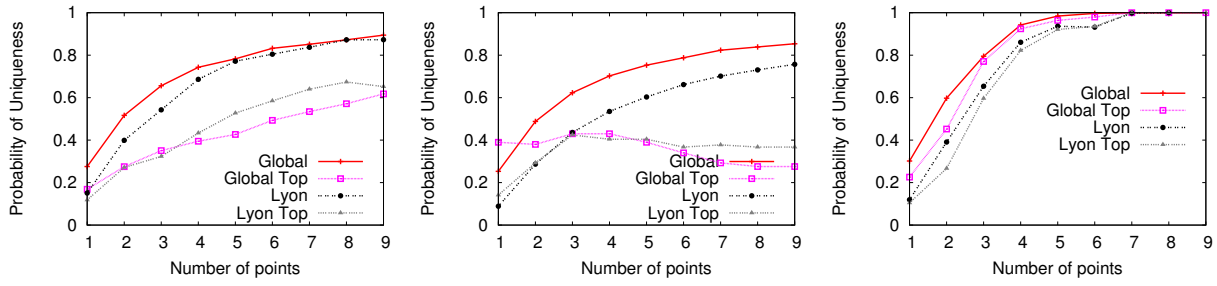
Figure 4: PRIVAMOV: Four spatio-temporal points are enough to uniquely identify 94% of the individuals. Focusing the analysis on restricted urban area reduces slightly the uniqueness.



(a) Global, GSM (b) Global, WiFi (c) Global, POI

(d) Lausanne, GSM (e) Lausanne, WiFi (f) Lausanne, POI

Figure 5: MDC: Four spatio-temporal points are enough to uniquely identify 94% of the individuals. Except for the WiFi-based mobility traces in the Lausanne area, exploiting the spatial information provides a better uniqueness.



(a) GSM (b) WiFi (c) POI

Figure 6: PRIVAMOV: Uniqueness assessment from the probabilistic model [3] gives an upper bound of the uniqueness compared to the deterministic approach [4].

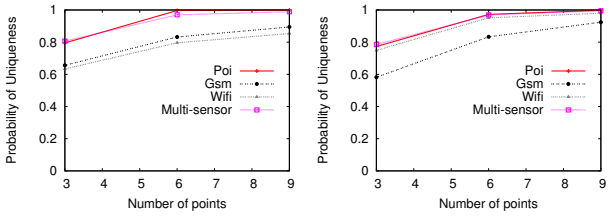
tion privacy-preserving mechanisms and the outcome of the previously defined de-anonymisation attack.

4.1 Quantifying Uniqueness

We quantify the uniqueness of mobility traces built from the GPS, WiFi, and GSM data collections. Figures 4 and 5 depict for the PRIVAMOV and the MDC datasets respectively the probability to be unique according to the number of points in the considered mobility trace. The evaluation reports results for spatial, temporal, and spatio-temporal traces, and for both the global dataset and a sub-area excluding the suburbs of the respective cities.

As shown in these figures, the results for spatio-temporal mobility traces from GSM, WiFi, and GPS report a strong uniqueness. More precisely, four spatio-temporal points are enough to identify between 89% and 99% of the users, depending on the mobility traces. This high uniqueness is the result of combining the temporal and the spatial mobility information of users, which are discriminative enough to uniquely identify them.

Furthermore, results show the high importance of the temporal information for uniquely identifying individuals. For instance, in the PRIVAMOV dataset, using the temporal information provides almost the same uniqueness as leveraging the spatial information for WiFi-based mobility traces and a better uniqueness for mobility traces extracted from the GSM data collection (83% against 74% with 4 points). These temporal traces reflect the time when a user moves or when he is inside a POI. Results show that this temporal information can also provide an important mobility footprint, which is sufficient to uniquely identify a large proportion of users. This temporal information is however not efficient for mobility traces extracted from POIs for the PRIVAMOV dataset. This means that the temporal information reflecting when a user moves provides a better fingerprint than the information reflecting when this user is inside a POI such as home or work place. Indeed, a user spends more time in a POI during the day than time moving, making this information more unique. For the MDC dataset, the temporal traces provides a 10% lower uniqueness than the spatial traces for GSM-based mobility traces while providing on average a 45% lower uniqueness for mobility traces extracted from WiFi and GPS data collections. Furthermore, results show that in the considered sub-areas, the temporal traces provide a better uniqueness than the spatial traces for mobility traces based on the GSM and the WiFi respectively on both datasets.



(a) PRIVAMOV (b) MDC

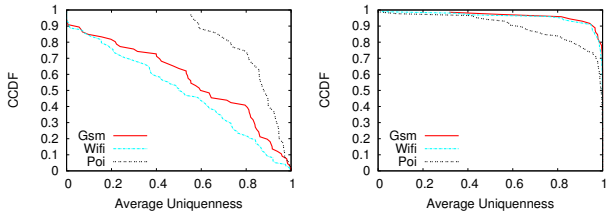
Figure 7: Using a multi-sensor fingerprints does not improve the uniqueness compared to using a fingerprint from one sensor.

Interesting enough, the uniqueness evaluation of mobility traces extracted from POIs for the PRIVAMOV dataset shows that the spatial only mobility patterns are more unique than the spatio-temporal ones (89% against 86% with 4 points). This is mainly due to the particularly low uniqueness provided by the temporal information in mobility traces from POIs in this dataset.

Moreover, results for the MDC dataset report a slightly smaller uniqueness than in the PRIVAMOV dataset. In addition, the uniqueness of spatial only mobility traces from the Lausanne area reflects that an anonymisation mechanism has been applied on the raw data before the release of the dataset (shown also by the shape of the CCDF Figure 3g where most of the users have been associated to many different WiFi access points).

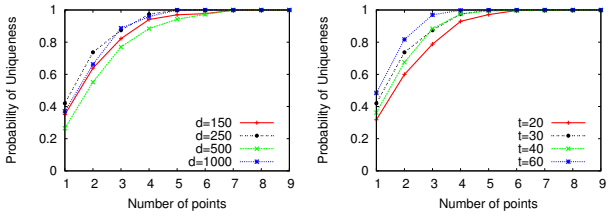
Additionally, we show that considering a sub-area slightly reduces the uniqueness. Indeed, focusing on the mobility of users in a limited and dense urban sub-area excludes isolated points that could be unique to users and may lead to easily identify them (e.g. a weekend in a family home).

We also evaluate the uniqueness of multi-sensor mobility traces instead of considering only information from the same sensor. Specifically, we build multi-sensor spatial only mobility traces mixing information from the GPS, the WiFi, and the GSM sensors. These multi-sensor mobility traces gather the same quantity of information from each data collection, as a consequence the number of points of these traces is a multiple of three. Figure 7 com-



(a) Spatial (b) Spatio-Temporal

Figure 8: PRIVAMOV: the uniqueness of users from spatial mobility traces is highly heterogeneous over users while the temporal information drastically increases the uniqueness compared to only considering the spatial information.



(a) Spatial Resolution (b) Temporal Resolution

Figure 9: MDC - POI extraction: Only the temporal resolution used to build POI is directly correlated to the uniqueness.

pares for both datasets the probability of uniqueness of these multi-sensor mobility traces against mobility traces from the GPS, the WiFi, and the GSM sensor only with the same number of points (i.e. multiple of three). Results show that mixing information from each sensor does not improve the uniqueness compared to the data collection which provides the best uniqueness (i.e., the POI-based mobility traces in our case). However, using information from multiple sensors avoids to know a priori which sensor provides the best uniqueness.

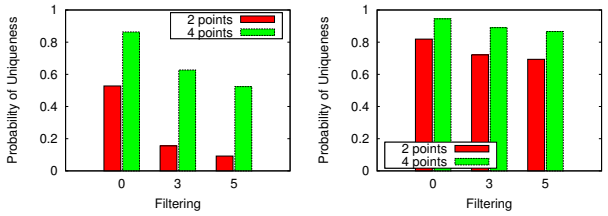
Finally, we compared the uniqueness measured by the probabilistic model proposed in [3] and the deterministic model proposed in [4]. As described in Section 3.3, the former uses random mobility traces while the letter uses mobility traces composed of the top n points the most frequently visited by users. Figure 6 depicts for the PRIVAMOV dataset the probability to be unique from both models for mobility traces from the GSM, WiFi, and POIs. Results show that the uniqueness measured from the model using the most visited spatial points (i.e., Global Top and Lyon Top) is the lowest (up to 40% less). Indeed, as shown in the CCDF Figure 3g, a majority of WiFi access points, POIs, and GSM antenna have been uniquely visited by one user. As a consequence, random-based mobility traces are more likely to include points with a strong uniqueness than mobility traces based on the most frequently visited points.

4.2 Variability of the Uniqueness

Depending on their mobility traces, all users do not have the same level of uniqueness. To highlight the variability of the uniqueness over the population of users, Figure 8 depicts the CCDF of the average uniqueness of users. Results show that the uniqueness of users from the GSM and the WiFi spatial only data collections are almost uniformly distributed in $[0 : 1]$ while this uniqueness is stronger and distributed only on $[0.5 : 1]$ for mobility traces built from POIs. These distributions reflects the large variability of the uniqueness of the spatial only mobility traces of users. Contrastingly, uniqueness from spatio-temporal information has less variability where most of the users are highly unique. This result also clearly demonstrates that the temporal dimension drastically increases the uniqueness of users compared to only considering spatial information.

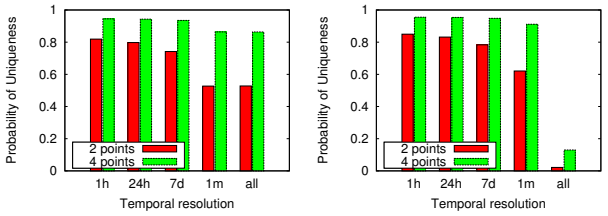
4.3 Impact of POIs Extraction Parameters

The POIs extraction performed from the GPS data collection has an impact on the POI-based mobility traces as shown in Figure 1, especially the considered diameter and the stay time. Shortly, a longer considered stay time reduces the number of identified POIs by filtering locations where the user did not stay long enough, while a larger diameter reduces also the number of identified POIs by aggregating small POIs in larger POIs. To assess the impact of these parameters, Figure 9 depicts the uniqueness according to both a varying diameter and different values of stay time. Results show that the temporal resolution



(a) Spatial (b) Spatio-Temporal

Figure 10: MDC - WiFi: A naive spatial filtering scheme removing too unique spatial information drastically decreases the uniqueness.



(a) Global (b) Lausanne

Figure 11: MDC - WiFi: Reducing the temporal resolution only slightly decreases the uniqueness.

(i.e., parameter t) used to extract POIs has a direct correlation to the uniqueness (Figure 9b), a larger t increases the uniqueness. Indeed, a larger t reduces the number of POIs in the resulting mobility traces, and consequently decreases the probability to find other users sharing the same POIs. However, while the spatial resolution (i.e., parameter d) impacts the uniqueness, there is no direct correlation with d (Figure 9a).

4.4 Impact of Spatial Filtering

As shown by the CCDF in Section 3.2, many GSM antennas, WiFi access points and POIs have been visited only by one user. Obviously, the unique nature of this information leads to improve the capacity to uniquely identify users. To quantify the impact of spatial components (i.e., GSM antennas, WiFi access points, or POIs) only visited by few users, we evaluate the uniqueness when these points are filtered from the mobility traces of users. Figure 10 reports the uniqueness of WiFi-based mobility traces of 2 and 4 spatial and spatio-temporal points from the MDC dataset when spatial points only visited by at most 3 and 5 users have been removed. Results show that this naive spatial filtering scheme drastically reduces the uniqueness for spatial mobility traces. Indeed, this naive filtering scheme tends to remove the too unique spatial information. However, as this filtering scheme only removes spatial information, the impact on the uniqueness for spatio-temporal mobility traces is less important.

4.5 Impact of Temporal Cloaking

Temporal cloaking reduces the temporal resolution of the mobility traces. Reducing the temporal resolution means aggregating spatial information in larger time units and thus reducing the information provided by the mobility traces. As a consequence, with a less precise timescale, users are more likely to share common spatial points with others at the same time slot. Figure 11 reports the uniqueness of mobility traces of 2 and 4 spatio-temporal points based on the WiFi data collection of the MDC dataset with a varying temporal resolution, from 1 hour to 1 month. This figures also reports the uniqueness when the temporal resolution includes the whole dataset (noted all in the figure) which means that all temporal information are removed and only the spatial information is used. Predictably, results show that a smaller temporal resolution reduces the uniqueness. This decrease is however very small. Indeed, the uniqueness only slightly reduces from 1 hour to 7 days regardless the number of spatio-temporals points. As the uniqueness of spatio-temporal mobility traces leverages information from both the temporal and the spatial information of users, reducing only the temporal information is not enough to counterbalance the spatial information. This observation supports results of previous studies [4, 3, 13, 2] showing that reducing the uniqueness through generalization requires a very coarse-grained information which drastically reduces the utility of the protected data.

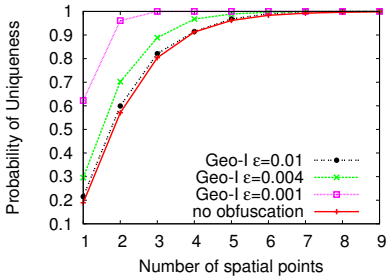


Figure 12: MDC - POI: The impact of GEO-I on the POI extraction on obfuscated data (i.e., the noise injection drastically reduces the number of POI) affects the uniqueness, the more noisy, the more unique.

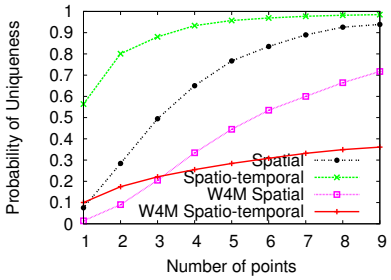


Figure 13: PRIVAMOV - POI: The impact of W4M on the POI extraction affects the uniqueness, the more noisy, the more unique.

4.6 Impact of GEO-I

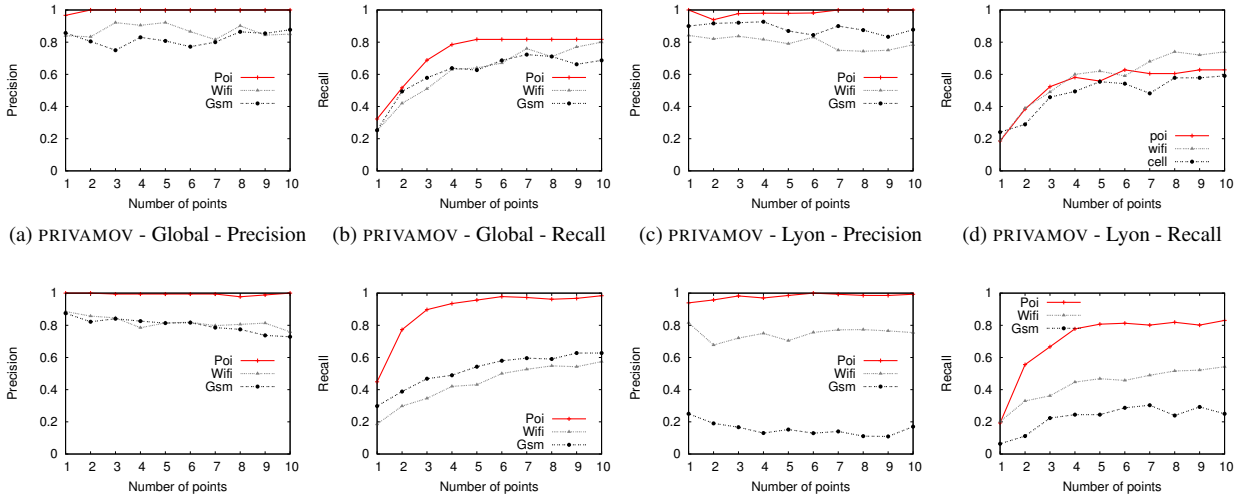
As described in Section 3.5, GEO-I obfuscates the spatial information of users with noise injection in the GPS coordinates. The level of noise is controlled by an ϵ parameter, the smaller ϵ, the noisier the obfuscation. This noise applied on the original dataset drastically impacts the POI extraction in the obfuscated data. Indeed, this noise injected in the GPS coordinates reduces the probability to have long enough data points in the same diameter to identify a POI. In the MDC dataset for instance, users have on average 28.9 POIs with an $\epsilon = 0.01$ while this number of POIs decreases to 19.8 and 4.7 with ϵ values equal to 0.004 and 0.001, respectively. Figure 12 shows the uniqueness of the spatial information according to different values of ϵ for the MDC dataset. Results show that the more obfuscated the mobility traces, the more unique they are. This result is mainly due to the decreasing number of POIs in the mobility traces when the obfuscation is enhanced, and thus reduces the probability to find other users sharing the same POIs. Interesting enough, using GEO-I with a low obfuscation (i.e., $\epsilon = 0.01$) provides almost the same uniqueness as exploiting the original mobility traces.

4.7 Impact of W4M

As described in Section 3.5, W4M obfuscates the spatial and the temporal information to provide k-anonymity by ensuring that k users are within the same cylinder. Contrary to GEO-I, the spatial and temporal data manipulation of W4M does not impact the POI extraction. In the considered PRIVAMOV dataset, users have on average 19.9 POIs without obfuscation while this number of POIs increases to 21.7 when W4M is applied on the mobility traces. Figure 13 depicts the probability of uniqueness of both the spatial and spatio-temporal data collections with and without obfuscation from W4M for the PRIVAMOV dataset. Results show in both cases (i.e., for spatial and spatio-temporal data collections) that applying this obfuscation scheme greatly decreases the uniqueness (down to 32% and 67% for the spatial and the spatio-temporal, respectively). This decrease is a direct result of the k-anonymity scheme of W4M which aims to avoid uniqueness by ensuring that each POI is shared by at least k users.

4.8 De-anonymisation attack

As described Section 3.4, we implemented a de-anonymisation attack to re-identity users from a training to a testing set built from the spatial only data collection. Figure 14 depicts for both datasets the precision and the recall for a varying size of background knowledge preserved for each user from the training set (i.e., the number of identifiers the most visited). Results show that the de-anonymisation attack can re-identify users with a high success rate. Using as background knowledge the most visited POIs provides up to almost a perfect precision for 80% and 98% of the users for PRIVAMOV and MDC dataset, respectively. Using background information



(a) PRIVAMOV - Global - Precision (b) PRIVAMOV - Global - Recall (c) PRIVAMOV - Lyon - Precision (d) PRIVAMOV - Lyon - Recall

(e) MDC - Global - Precision (f) MDC - Global - Recall (g) MDC - Lausanne - Precision (h) MDC - Lausanne - Recall

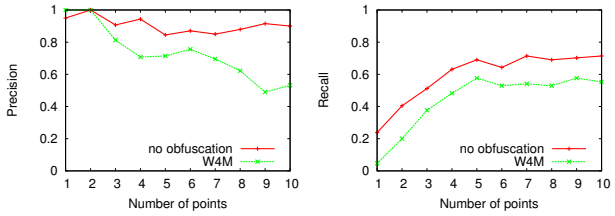
Figure 14: De-anonymisation attack re-identifies a large population of users with a high precision.

from the WiFi and the GSM reduces the precision by 20% on average and the recall by 10% and 30% for PRIVAMOV and MDC dataset, respectively. While the precision tends to decrease according to the size of the background knowledge, the recall inversely tends to increase. Indeed, as the size of the background information increases, identifying users from their mobility habits from the training set become easier (i.e., a better recall) but the fault positives also increase (i.e., a smaller precision).

Similar results are observed when the attack is performed on geolocated data restricted to a city (Figures 14c, 14d, 14g and 14h). Compared to using the data in the whole dataset, the precision provided by the WiFi and GSM data collections are slightly better while the recall is 20% lower. Interesting enough, for the MDC dataset restricted to Lausanne, both the precision and the recall provided by the GSM information are very low (i.e., around 20% regardless the size of the background knowledge). These low values are certainly due to the privacy preserving scheme applied on the raw data before releasing the dataset.

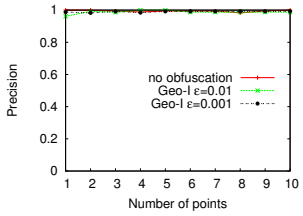
These results are consistent with the observations done through the uniqueness assessment of Section 4.1: the uniqueness drives the effectiveness of the de-anonymisation attack (i.e., the more unique, the more effective is the attack). Indeed, the POI-based spatial only mobility traces are the more sensitive to the de-anonymisation attack compared to the ones from the WiFi and GSM data collections (Figure 14), which also provide the stronger uniqueness (Figures 4 and 5). Moreover, the de-anonymisation attack is slightly less effective for the MDC dataset than for the PRIVAMOV dataset. Similarly, the MDC dataset reports a slightly smaller uniqueness than in the PRIVAMOV dataset. Lastly, considering smaller and denser sub-area (resp.) reduces the effectiveness of the attack, and reduces the uniqueness (resp.).

To fully understand the impact of LPPMs, we also conduct the de-anonymisation attack on the obfuscated data. Figure 15 reports the precision and the recall of the de-anonymisation attack for the PRIVAMOV dataset performed on both un-obfuscated and obfuscated data from GEO-I and W4M. Results for W4M (Figure 15a and 15b) show that excepting for $n < 3$ where the precision of obfuscated data is higher than the un-ofuscated data, the precision tends to decrease according to n, the number of most visited points. However, slightly more users are identified (i.e., a better recall) with un-obfuscated data. Results for GEO-I (Figure 15c and 15d), show that the pre-

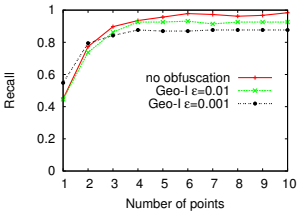


(a) PRIVAMOV - Precision

(b) PRIVAMOV - Recall



(c) MDC - Precision



(d) MDC - Recall

Figure 15: GEO-I and W4M fail to protect users against a de-anonymisation attack.

cision is similar with and without obfuscation and close to 1. The recall, in turn, slightly decreases according to the level of the obfuscation but remains high (i.e., > 0.8 from $n = 2$).

5 Discussions and conclusions

This paper reports an extensive experimental evaluation of the uniqueness of mobility traces of users collected through different sensors (i.e., GPS, GSM, and WiFi). We show that these mobilty traces are highly unique: on average only four statio-temporal points are enough to uniquely identify on average 94% of the individuals. Moreover, we assess the uniqueness of the spatial-only (i.e., where users move) and the temporal-only (i.e., time when users move or are in POIs) information related to the mobility traces. Although the temporal information does not identify users as much as the spatial information on average, results show that the temporal footprint is enough to uniquely identify a large proportion of users and drastically improve the uniqueness when it is combined with the spatial information. We also highlight the heterogeneous nature of the uniqueness over the population of users as the mobility traces of some users are more unique than others. In addition, we compared two different models to quantify the uniqueness and show that the probabilistic model proposed in [3] provides an upper bound compared to the deterministic model proposed in [4]. Furthermore, similarly to previous studies, we show that reducing the uniqueness through generalisation requires a very coarse-grained information which drastically limits the utility of the data. Lastly, we show that a de-anonymisation attack re-identifying anonymous users using only a limited background information on the mobility habit of users provides a very high success rate, even if the raw data are obfuscated by classical LPPMs.

Besides, we evaluated the impact of different LPPMs on the uniqueness and show that obfuscation of GPS coordinates from GEO-I and W4M leads to opposite results. The obfuscation from GEO-I tends to decreases the number of extracted POIs. This side effect of fewer extracted POIs results in increasing the uniqueness. In contrast, the obfuscation scheme of W4M based on k-anonymity meets the expectations by greatly reducing the uniqueness.

We hope that the observations done in this paper can lead to the development of more effective LPPMs in the future. In particular, as most of the existing LPPMs are static (i.e., they apply the same level of obfuscation to all users) and focus their obfuscation on the spatial dimension, possible improvements may be the development of an adaptive LPPM that dynamically adapts the obfuscation simultaneously time and space with the respect of the considered user.

References

[1] H. Henttu, J.-M. Izaret, and D. Potere, “Geospatial Services: A $1.6 Trillion Growth Engine for the U.S. Economy,” http://www.bcg.com/documents/file109372.pdf, 2012.

[2] M. Gramaglia and M. Fiore, “Hiding Mobile Traffic Fingerprints with GLOVE,” in *CoNEXT*, Dec. 2015.

[3] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, 2013.

[4] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *MobiCom*, 2011, pp. 145–156.

[5] J. Mayer, P. Mutchler, and J. C. Mitchell, "Evaluating the privacy properties of telephone metadata," *Proceedings of the National Academy of Sciences*, vol. 113, no. 20, pp. 5536–5541, 2016.

[6] A. Cecaj, M. Mamei, and N. Bicocchi, "Re-identification of anonymized cdr datasets using social network data," in *PERCOM Workshop*, 2014, pp. 237–242.

[7] A. N. and V. S., "Robust de-anonymization of large sparse datasets," in *SP*, 2008, pp. 111–125.

[8] S. Gambs, M.-O. Killijian, and M. Núñez Del Prado Cortez, "De-anonymization attack on geolocated data," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, Dec. 2014.

[9] K. Andreas, G. Hugo, B. Tobias, R. Konrad, and F. Felix, "Fingerprinting mobile devices using personalized configurations." *PoPETS*, vol. 2016, no. 1, pp. 4–19, 2016.

[10] P. Eckersley, "How unique is your web browser?" in *PETS'10*, 2010, pp. 1–18.

[11] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting." *PoPETs*, vol. 2016, no. 1, pp. 34–50, 2016.

[12] O. Rebekah and G. Rachel, "Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution." *PoPETS*, vol. 2016, no. 3, pp. 155–171, 2016.

[13] M. Gramaglia and M. Fiore, "On the anonymizability of mobile traffic datasets," *CoRR*, vol. abs/1501.00100, 2014.

[14] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," in *SDM*, 2005, pp. 185–199.

[15] G. Acs and C. Castelluccia, "A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris," in *SIGKDD*, 2014, pp. 1679–1688.

[16] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential Privacy for Location-based Systems," in *SIGSAC*, 2013, pp. 901–914.

[17] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie, "Time Distortion Anonymization for the Publication of Mobility Data with High Utility," in *TrustCom*, 2015.

[18] K. Fawaz and K. G. Shin, "Location privacy protection for smartphone users," in *SIGSAC*, 2014.

[19] L. Sweeney, "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[20] C. Dwork, "Differential Privacy: A Survey of Results," in *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*. Springer-Verlag, 2008, pp. 1–19.

[21] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The New Casper: Query Processing for Location Services Without Compromising Privacy," in *VLDB*, 2006, pp. 763–774.

[22] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems," in *WWW*, 2007, pp. 371–380.

[23] L. S., "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, Oct. 2002.

[24] C. C. Aggarwal and P. S. Yu, *A Condensation Approach to Privacy Preserving Data Mining*, 2004, pp. 183–199.

[25] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *ICDE*, 2008, pp. 376–385.

[26] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie, "Adaptive Location Privacy with ALP," in *SRDS*, 2016.

[27] J. Freudiger, R. Shokri, and J.-P. Hubaux, "Evaluating the privacy risk of location-based services," in *FC*, 2012, pp. 31–46.

[28] "Priva'mov project: http://liris.cnrs.fr/privamov/project/."

[29] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen, "From big smartphone data to worldwide research: The mobile data challenge," *Pervasive Mob. Comput.*, vol. 9, no. 6, pp. 752–771, Dec. 2013.

[30] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and L. J., "Towards rich mobile phone datasets: Lausanne data collection campaign," in *ICPS*, 2010.

[31] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw gps data for geographic applications on the web," in *WWW*, 2008, pp. 247–256.

[32] "Location guard," Available online at https://github.com/chatziko/location-guard.

[33] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Information Systems Journal*, vol. 35, no. 8, pp. 884–910, Dec. 2010. [Online]. Available: http://www-kdd.isti.cnr.it/W4M/