SwapMob: Swapping trajectories for mobility anonymization

Julián Salas[1], David Megías[1], and Vicenç Torra[2]

[1] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC),
Barcelona, Spain.
CYBERCAT-Center for Cybersecurity Research of Catalonia.
jsalaspi@uoc.edu, dmegias@uoc.edu,
[2] School of Informatics, University of Skövde, Skövde, Sweden.
vtorra@his.se.

Abstract. Mobility data mining can improve decision making, from planning transports in metropolitan areas to localizing services in towns. However, unrestricted access to such data may reveal sensible locations and pose safety risks if the data is associated to a specific moving individual. This is one of the many reasons to consider trajectory anonymization.

Some anonymization methods rely on grouping individual registers on a database and publishing summaries in such a way that individual information is protected inside the group. Other approaches consist of adding noise, such as differential privacy, in a way that the presence of an individual cannot be inferred from the data.

In this paper, we present a perturbative anonymization method based on swapping segments for trajectory data (SwapMob). It preserves the aggregate information of the spatial database and at the same time, provides anonymity to the individuals.

We have performed tests on a set of GPS trajectories of 10,357 taxis during the period of Feb. 2 to Feb. 8, 2008, within Beijing. We show that home addresses and POIs of specific individuals cannot be inferred after anonymizing them with SwapMob, and remark that the aggregate mobility data is preserved without changes, such as the average length of trajectories or the number of cars and their directions on any given zone at a specific time.

0.1 ...TO DO*

Given a probability model, such as Markov chain models, of the data, and a subsequent decision problem (eg. traffic flow prediction from mobility siumlations based on the learnt Markov chain model), we can limit ourselves to data-anonymizing operators that preserve the (minimal) sufficient statistic of the model.

Intuitively speaking, all the information in the data about the parameters of the model are in the sufficient statistic. Therefore, by anonymizing the data while preserving the sufficient statistics is decision-theoretically optimal, in the sense of maximizing utility from optimal estimates of the parameters in the probability model.

A *sanitizer* $\mathcal{S}$ is a map from the data space $\mathbb{D}$ to a sanitized data space $\mathbb{D}'$, i.e., $\mathcal{S}(d) : \mathbb{D} \to \mathbb{D}'$. A sanitizer may or may not preserve the minimimal sufficient statistics in the data. A *sufficient sanitizer* preserve the minimal sufficient statistics.

... then we can choose between them based only on privacy concerns.

- This is my interpretation of the notes from last day :

We consider probabilistic models for a database D and statistical inference by time homogeneous discrete time Markov Chains with probability matrix P and rate matrix Q which is a sufficient statistic of D for this model

- Where are these used in transportation (some citations and brief explanations)
- Rigth now I could generate the probability matrices with states S_i as spatial points with rounded coordinates (coarsened locations), it is quite straightforward, and they still represent square areas in the map.
- Prove that SwapMob preserves the minimal sufficient statistic
- Argue why other methods would not preserve the minimal sufficient statistic
- Explain synthetic generation of trajectories

For future work:

- Calculating privacy and utility may be left as future work.

1 Introduction

With the pervasive use of smartphones and the location techniques such as GPS, GSM and RFID, the opportunities to deliver content depending on current user location have increased. Location Based Services (LBS) provide considerable advantages such as allowing users to benefit from live location-based information for transportation, recommendations of places of interest, or even the opportunity to meet friends in nearby locations. Such location-based data can be useful also for intelligent transportation systems, in which vehicles may serve as sensors for collecting information about traffic jams, weather, and road conditions.

However, revealing users' locations may have some privacy risks. If the data is linked to the real identities it may reveal personal preferences (e.g., sexual, political or religious orientation), or it may be used for inferring habits and know the time when a person is at home or away. To avoid such inconveniences, a variety of anonymization techniques have been developed to hide the identity of the user or her exact location, e.g, [26].

Moreover, as Giannotti et al. mention in [10], big data (in particular trajectory data) may be used to understand human behavior through the discovery of individual social profiles, by the analysis of collective behaviors, spreading epidemics, social contagion, and to study the evolution of sentiment and opinion; however, trusted networks and privacy-aware social mining must be pursued and methods for protection and anonymization for such data must be developed to enforce the data subjects' rights and promote their participation.

2 Related work

Different solutions have been proposed for anonymizing trajectories in data publishing. Abul et al. [1], propose the (k, δ)-anonymity model, which consists on publishing a cylindrical volume of radius δ that contains the trajectory of at least k moving objects. Note that this idea is an extension of the concept of k-anonymity for databases [22].

Terrovitis and Mamoulis [27] consider a discrete spatial domain, e.g., spatial information is given in terms of addresses in a city map. Hence, the user trajectories are expressed as sequences of POIs. They present the use case of the RFID cards from the Octopus[3] company in Hong Kong, which collects the transaction history of its customers. The company may want to publish sequences of transactions by the same person as trajectories, for extracting movement and behavioral patterns. However, if a given user, Alice, uses her card to pay at different convenience stores that belong to the same chain (e.g., convenience stores), that company may reidentify Alice if her sequence of purchases is unique in the published trajectory database.

A similar approach in [18] is obtained by transforming sequences by adding, deleting, or substituting some points of the trajectory, while preserving also frequent sequential patterns [2] obtained by mining the anonymized data.

In [14] and [13], Hoh et al. discuss the use of mobility data for transportation planning and traffic monitoring applications to provide drivers with feedback on road and traffic conditions. For modelling the threats to privacy in such datasets, they assume that an adversary does not have information about which subset of samples belongs to a single user, however by using multi-target tracking algorithms [19] subsequent location samples may be linked to an individual that is periodically reporting his anonymized location information.

In [13] they consider the attack of deducing home locations of users by leveraging clustering heuristics used together with the decrease of speed reported by GPS sensors. Then, propose data suppression techniques by changing the sampling rate (e.g, from 1 minute to 2,4 and 10) for protecting from such inferences.

In [14], in order to prevent adversaries from tracking complete individual paths, they propose an algorithm that perturbs slightly the trajectories of different individuals (to make them closer) in such a way that the adversary may not be able to follow which segment of the path corresponds to which user by using multi-target tracking algorithms. This is done with a constraint on the Quality of Service, which is expressed as the mean location error between the actual and the observed locations. They argue that adequate levels of privacy can only be obtained if the density of users is sufficiently high.

This is closely related to [3] in which Mix Zones are introduced, these are spatial areas on which users' location is not accessible, hence when users are simultaneously present on a mix zone, their pseudonyms are changed. This procedure is performed to difficult the linkage of the incoming and outgoing path segments to the same specific user.

[3] http://www.octopuscards.com/

They design a model for location privacy protection that aims to preserve the advantages of location aware services while hiding their identities from the applications that receive the users' locations. The existence of a trusted middleware system (or sensing infrastructure) is assumed and the applications register their interest in a geographic space with the middleware, such space is called application zone. Examples of such application zones are hospitals, universities or supermarket complexes, in general it could be any open or closed space.

The regions in which applications cannot trace user movements are called mix zones, and the borders between a mix zone and an application zone are called boundary lines. Applications do not receive traceable user identities, they receive pseudonyms that allow communication between them. Such communication passes through the trusted intermediary and the pseudonyms of users change when they enter a mixed zone.

In order to measure location privacy, Beresford and Stajano [4] define the anonymity set as the group of people visiting the mix zone during the same time period. However, as the boundary and time when a user exits a mix zone is strongly correlated to the boundary and time when the user enters it, such information may be exploited by an attacker, therefore they use the information theoretic metric that Serjantov and Danezis [24] proposed for anonymous communications which considers the varying probabilities of users sending and receiving messages through a network of mix nodes.

This is modeled in [4] as a movement matrix in which they record the frequency of ingress and egress points to the mix zone at several times. Then, a bipartite weighted graph is defined in which vertices model ingress and egress pseudonyms and edge weights model the probability that two pseudonyms represent the same underlying person. Therefore, a maximal cost perfect matching of these graphs represents the most probable mapping among incoming and outgoing pseudonyms.

However, since the solution to many restricted matching problems (such as this one) is NP-hard [25], Beresford and Stajano [4] describe a method for achieving partial solutions.

An approach that does not consider middleware to obtain location privacy is proposed in Chapter 9 from [11]. It consists of a system with an untrusted server and clients communicating in a P2P network for privacy preserving trajectory collection. The aim of their data collection solution is to preserve anonymity in any set of data being stored, transmitted or collected in the system. This is achieved by means of k-anonymization and swapping. Briefly, the protocol consists of the clients recording their private trajectories, cloaking them among k similar trajectories and exchanging parts of those trajectories with other clients in the P2P network. However, the final step (the data reporting stage) clients send anonymous partial trajectories to the server, that have been generated in such a way that the server can filter all the synthetic trajectory data that has been generated for cloaking during the process, and recover the original trajectory.

One of the advantages of performing trajectory anonymization on the user side, as in [20], is that the anonymization process is no longer centralized. Thus data subjects gain control, transparency and more security for their data.

For a brief overview of privacy protection techniques and a discussion of k-anoymity and differential privacy models in different frameworks, cf. [21].

In [7], a differential privacy model for transit data publication is considered, using data from the Société de Transport de Montréal (STM). The data are modeled as sequential data in a prefix tree that represents all the sequences by grouping the sequences with the same prefix into the same branch. Their algorithm takes a raw sequential dataset D, a privacy budget ϵ, a user specified height of the prefix tree h and a location taxonomy tree T, and returns a sanitized dataset $\tilde{\mathcal{D}}$ satisfying ϵ-differential privacy. For measuring utility, in the STM case, sanitized data are mainly used to perform two data mining tasks, count query and frequent sequential pattern mining [2].

Other ϵ-differentially private mechanism for publishing trajectories called SDD (Sampling Distance and Direction) can be found in [15]. They focus on ship trajectories with known starting point and terminal point. And consider that two trajectories T and T' with the same number of positions are adjacent if they differ at exactly one position excluding the starting point and the terminal point.

In [28], a differentially private algorithm for location privacy is proposed, following a discussion on the (in)applicability of differential privacy in a variety of settings, such as [6] and [16]. Their algorithm considers temporal correlations modeled as a Markov chain and proposes the "δ-location set" to include all probable locations (where the user might appear). The authors argue that, to protect the true location, it is enough to hide it in the δ-location set in which any pairs of locations are not distinguishable. However, they leave the problem of protecting the entire trace of released locations as future work.

In this paper, we present an anonymization method considering that the data are dynamic, the rate at which the information is collected is not constant, and the databases are being generated as the data is received.

3 Proposed method: SwapMob

We propose a method for anonymization of mobility data by swapping trajectories, which works in a similar way as the mix zones but in a non-restricted space.

Our algorithm (SwapMob) simulates an online P2P system for exchanging segments of trajectories. That is, when two users are near they interchange their partial trajectories, see section 3.1. In this way, all users' trajectories are mixed incrementally, and the moving users keep generating segments of trajectories that are being swapped. In the end, each trajectory retrieved is made of small segments of trajectories of different individuals, who have met during the day, as depicted in Figure 1. Hence, the relation between data subjects and their data is obfuscated while keeping a precise aggregated data, such as the number of users in each place at each time and the locations that have been visited by different anonymous users.

We formalize our method after a brief explanation of previous definitions and assumptions.

3.1 Definitions

We assume that we have a database in which the i-th observation is a tuple (ID_i, lat_i, long_i, t_i) that consists of the individual's identifier (ID_i), the latitude (lat_i), longitude (long_i) and timestamp (t_i).

Then, the trajectory T_x of an individual x will consist of all the observations with identifier x ordered by their timestamps t_i. These can be represented as $T_x = (x_1, x_2, \ldots, x_m)$ if there are m observations for individual x.

We say that *two individuals meet* or their trajectories cross (on points x_i and y_j) if they have been co-located. We denote this by $x_i \approx y_j$. Note that being co-located depends on thresholds for proximity (χ) and time (τ), since the sampling rate of positions is not regular nor constant. Moreover two persons cannot be in the exact same place at the same time.

We define a *matching* as a maximal subset of pairs of elements of a set.

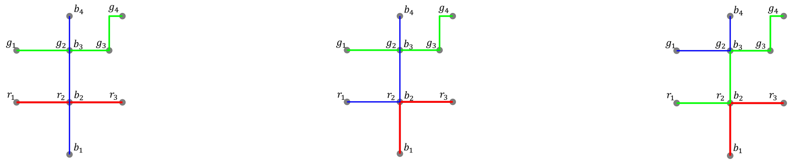
We denote by $Sw(T)$ the resulting trajectory after all swaps have been applied to T. Next, we define the following two primitives for our algorithm: *generate random matching* and *swap*.

1. *Swap:* Given two trajectories $T_x = (x_1, \ldots, x_i, x_{i+1}, \ldots)$ and $T_y = (y_1, \ldots, y_j, y_{j+1}, \ldots)$ that meet in points x_i and y_j, a swap of T_x with T_y at points x_i and y_j results in $Sw(T_x) = (y_1, \ldots, y_j, x_{i+1}, \ldots)$ and $Sw(T_y) = (x_1, \ldots, x_i, y_{j+1}, \ldots)$.
2. *Generate random matching:*
   Given a set of elements $S = s_1, s_2, \ldots, s_m$, we generate a random matching by making pairs of the first $m/2$ with the following $m/2$ numbers, followed by a random permutation of all numbers m.

Note that, in case that the number of elements m is odd, to generate a matching we must leave out one element and that all possible random matchings can be generated following our procedure.

Crossing paths and Swapping We propose a model such that two peers get in contact (meet) if they have been co-located on a similar timestamp depending on parameters of proximity χ and time τ.

Next, we simulate SwapMob protocol by swapping the users IDs when the users have passed close enough. We calculate the set of users that get in contact in a given time interval, and choose a random matching among them when they are even and a matching of all but one, when they are odd. Here, the swapping is carried out in a pairwise manner, but it could be done as a permutation such as in [4].



(a) Original trajectories (b) After first swap (c) After second swap

Fig. 1: Three trajectories before and after swapping

Algorithm 1: Offline algorithm for swapping trajectories

```
Input: Trajectory Database. Thresholds for time $\tau$ and proximity $\chi$.
Output: Swapped trajectories identifiers $Sw(T_i)$.
Partition the timestamps $t = \bigcup \tau_j$ in intervals of length $\tau$
for each pair of registers $i, j$ in interval $\tau_j$ do
    if $dist(l_i, l_j) < \chi$ then
        add $i, j$ to close records list (possible swaps) $S_{\tau_j}$ at the given time
          interval.
    end
end
generate random matching with possible swaps in $S_{\tau_j}$
order all swaps in $\bigcup S_{\tau_j}$ by timestamp
for each pair $i \approx j$ in $\bigcup S_{\tau_j}$ do
    swap $T_i$ with $T_j$
end
return Swapped trajectories $Sw(T_i)$
```

Note that changing pseudonyms (IDs) is equivalent to swapping the partial trajectories.

In Figure 1, we present an example of three simple trajectories crossing T_r, T_g, T_b. We assume that they are moving from left to right and upwards, $T_r = (r_1, r_2, r_3)$, $T_g = (g_1, g_2, g_3, g_4)$ and $T_b = (b_1, b_2, b_3, b_4)$. Note that we are also assuming that the blue trajectory meets the red trajectory first ($b_2 \approx r_2$) and then the green trajectory ($b_3 \approx g_2$). In this tiny example, we can see how the iterative swaps preserve parts of the trajectory intact, but at the end each trajectory has parts of many others, such as the green one which ends having a segment of the blue trajectory, a segment of the red and a segment of its original trajectory $Sw(T_g) = (r_1, r_2, b_3, g_3, g_4)$.

3.2 SwapMob anonymizer

We follow a similar architecture to the one in [13] in which a Trusted Third Party (TTP) knows the vehicles identities but can not access sensor information (such as position and speed); and a Service Provider (SP) knows the sensor measures but not the identities. Further, the SP calculates which records are close to each other without knowing to which individual they belong and communicates them to the TTP (in this case SwapMob anonymizer) such that it can swap their identities without knowing at which location they were.

This is achieved in the following way (See Figure 2):

1. Users communicate with SwapMob, sending their sensor data (M) encrypted with the public key (K_{SP}) of SP. SwapMob keeps the number of register (i), which user has sent it (u_i), its current pseudonym (ID_i), the timestamp (t_i) and the encrypted sensor data $E(M_i, K_{SP})$, which includes their encrypted location (l_i).
2. SwapMob sends the vector $(i, t_i, E(M_i, K_{SP}))$ to the SP, who decrypts $E(M_i, K_{SP})$ and keeps a buffer of data on interval τ_j that contains all timestamps between timestamp t_j and t_{j+1} and has length τ, that is $\tau_j = \{t : t_j < t < t_{j+1}\}$.
3. SP sends the set S_{τ_j} of registers that were at distance less than the predefined threshold χ during the interval of time τ_j back to SwapMob, more formally $S_{\tau_j} = \{i, i' : d(l_i, l_{i'}) < \chi \text{ and } t_i, t_{i'} \in \tau_j\}$. SwapMob calculates the swaps and stores the users and swapped IDs list, that is, for every record i SwapMob keeps the corresponding swapped id $Sw(ID_i)$ and the user (u_i) to which such pseudonym corresponds.
4. Finally, every given period of time which could be daily, weekly or monthly, SwapMob reports the list of $(i, Sw(ID_i))$ to SP.

The authentication data integrity of the communications can be guaranteed with a hash-based message authentication code.

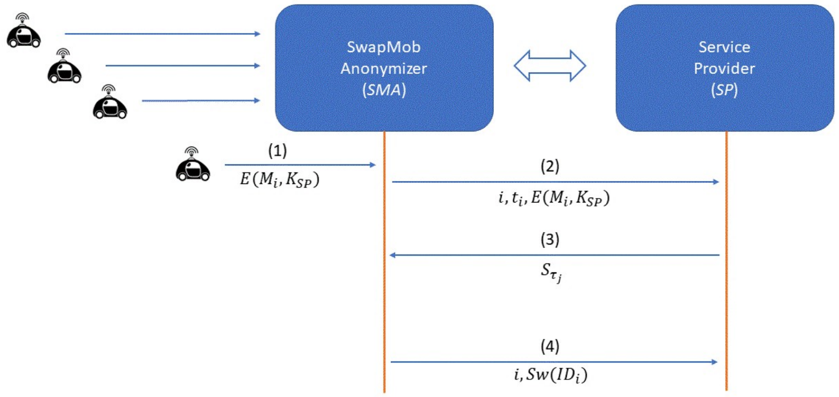


Fig. 2: Architecture of our system

In this way, SP obtains the measures of all sensors M in real-time (Step 2), and at the end of the day also gets the anonymized trajectories of the users that generated them (Step 4). Even, though SP knows which records belong to S_{τ_j} (Step 3), SP does not know to which other record they have been swapped during period τ_j, and by the iterative swappings it gets even harder to associate them to a specific user.

At the same time, SwapMob only knows the users, the timestamps at which they have crossed, and the reported trajectories are already anonymized by SwapMob (Step 4).

Our system, can be applied for the use case proposed in [3], by defining a set of swap zones (similar to the mix zones) and adding the restriction that the swapping cannot be performed outside such places. Then, the spatio-temporal trajectories of users between such swap zones could be monitored in an anonymous and precise way.

However, there will still be some differences. Namely, the swap zone that we consider is the entire application zone, whereas in [3] a user entering a mix zone can be distinguished from another user emerging from the same zone if the size of the mix zone is too large.

This same argument justifies that the distance and time parameters, χ and τ must not be too large either in our algorithm, otherwise swapping could not be credible.

3.3 Protecting against reidentification

It is well known that de-identification does not necessarily means anonymization. The same attributes that are used for extracting knowledge, may be used for pointing to a specific individual, and uniquely relating his/her data to her real identity.

Other notions of privacy are defined depending on the context, which may be of statistical databases [9], networks [33], or geo-located data.

By identifying the POIs of an individual, it is possible to infer his habits (e.g., does sport, travels a lot), the locations that he visits frequently (may be related to political or religious beliefs) or even related to health (clinics, hospitals). This may also be used to infer his schedule, predict his future locations, and learn his past locations and possibly his personal relations by observing frequent or periodic co-location. Moreover, such habits and locations can be easily used to reidentify the individuals behind the data. As it has been proven on previous anonymity studies on anonymity of home/work location.

Regarding this topic, Golle and Partridge studied in [12] workers who revealed their home and work location with noise or rounding on the order of a city block, a kilometer or tens of kilometers (census block, census tract, or county) and showed that the sizes of the anonymity set were respectively 1, 21 and 34,980. That is, when the data granularity was on the order of a census block, the individuals were uniquely identifiable, and for granularities on the order of census track or county, they were protected within sets of size 21 or 34,980. In [31], Zang and Bolot inferred the top N locations of a user from call records and correlated such information with publicly-available side information such as census data. Then, they showed that the top 2 locations likely correspond to home and work location and that the anonymity sets are drastically reduced if an attacker infers them.

Therefore, for protecting the individuals against reidentification, is crucial to protect their home addresses and POIs, to provide them with minimum guarantees of keeping them anonymous. Swapped data may not allow for following a specific individual and his whereabouts, and thus, this will not permit personalization or individual classification, which are ways of protecting their privacy.

A different approach regarding the possibility of reidentification and the (im)possibility of protection, is in [17], where they measure the uniqueness of human mobility traces depending on their resolution and the available outside information, assuming that an adversary knows p random spatio-temporal points. Then, they coarsen such data spatially and temporally to find a formula for uniqueness depending on such parameters.

We argue that SwapMob preserves anonymity by dissociating the segments of trajectories from the subject that generated them.

An attacker may know several spatio-temporal points of an individual that uniquely identify him. However, to link a register in the anonymized database to such an individual, the points known by the attacker must belong to the same trajectory after swapping. In most cases, the attacker will not learn the entire trajectory information since the published trajectory is made of segments from many different individuals. Of course, when publishing the trajectories, it should be noted that they have been generated by SwapMob, and the anonymization may be reversed if the SwapMob Anonymizer and the Service Provider (see Figure 2) share their information for guaranteeing accountability.

3.4 Utility of swapped data

In this paper we are assuming that the interest of using data anonymized by SwapMob is for making mobility maps and predictions that may be useful for intelligent transportation systems and for planning in a city. As Hoh and Gruteser proposed in [14], pre-specified vehicles could periodically send their locations, speeds, road temperatures, windshield wiper status and other information to the traffic monitoring facility. These statistics can provide information on the traffic jams, average travel time or the quality of specific roads, and can be used for traffic light scheduling and road design.

Furthermore, the sensors do not necessarily have to be attached to vehicles, they could be carried on mobile phones, and the utility of using the individuals for sensing is preserved, since all their sensor data, including all their movements and timestamps (in aggregate) are kept intact by SwapMob.

In [5], a real-time urban monitoring platform and its application to the City of Rome was presented, they used a wireless sensor network to acquire real-time traffic noise from different spots, GPS traces of locations from 43 taxis and 7268 buses, and voice and data traffic served by each of the base transceiver stations from a telecom company in the urban area of Rome. These are few examples of sensor that could be carried by individuals, anonymized and transmitted to a service provider via SwapMob.

Another example is the offline mining in [29] representing the knowledge from taxi-drivers as a landmark graph could be done with SwapMob anonymized data. A landmark is defined as a road segment that has been frequently traversed by taxis, and a directed edge connecting two landmarks represents the frequent transition of taxis between the two landmarks. This graph is then used for traffic prediction and for providing a personalized routing service.

In general, lossless maps of flows in the city can be obtained by using SwapMob at several aggregation levels and for different timestamps.

4 Empirical evaluation

We tested our algorithm on the T-drive dataset [30],[29] which contains the GPS trajectories of 10,357 taxis during the period of Feb. 2 to Feb. 8, 2008 within Beijing. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches to 9 million kilometers. It is important to note that not all taxis appear every day and not all report their positions at the same interval. The average sampling interval is about 177 seconds and 623 meters. Each line contains the following data: taxi id, date time, longitude, latitude.

4.1 Reidentification by POIs and home location

Recall that our main privacy motivations are to protect locations related to people's habits and also the association of the trajectories to specific individuals.

We show in the next subsections how to infer points of interest of an individual such as his home location. Then, we show that such locations cannot be inferred after applying SwapMob to the data.

Inferring points of interest and home location In [32], if an individual remains at less than 200 meters from a given point during at least 20 minutes, then it is considered a stay point (or point of interest) thus the two thresholds used for detecting them are time $\tau = 20$ and distance $\chi = 200$.

In general, for obtaining points of interest, we discretize the space in square cells of 111.32 m (or 0.001 decimal degrees) and count which are the most populated cells for each individual, considering that the most populated one should contain the home location.

This is similar to [8] that discretize the world into 25 by 25 km cells and define the home location as the average position of check-ins in the cell with the most check-ins, see also [23].

In [3], Beresford and Stajano tested this kind of attack on real location data from the Active Bat, consisting in following the trajectory of a pseudonym to a "home" location (in this case the user's desk in the office) and successfully deanonymize all the users by correlating where does any given pseudonym spend most of its time and who spends more time than anyone else at any given desk.

We validated this assumption empirically by looking at the distribution of timestamps for such set of locations, which is greater than 5% and increases consistently from 20 h until 6 h where it reaches its peak and decreases below 5% around 10 h, see Figure 3. Also, if we assume that those with more than 5% of relative frequency, are the correct home locations, we obtain that the correct home locations amount for 83,7% of the total, a very similar percentage to the one obtained in [8] by manual inspection which was 85% accuracy.

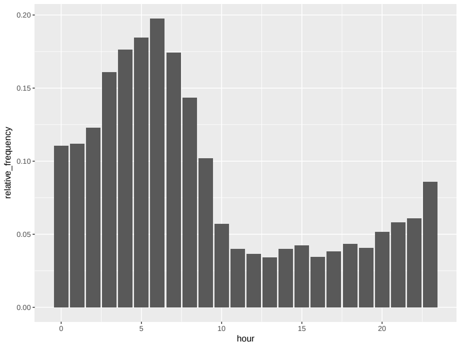


Fig. 3: Frequency histogram of hours at deduced home locations, as a percentage of total number of measurements for each hour.

Home locations after swapping We validated the results of our anonymization technique by trying to infer users home locations after swapping. Considering that the home location is the most relevant POI, we can argue that these experiments prove that not only home locations will be protected by our method, but all POIs.

For carrying out the swapping process, we assumed that two taxis were colocated if they passed at distance at most 111 meters approximately ($\chi = 0.001$) in a 1-minute interval ($\tau = 60$). Note that this is about 6 and 3 times less the average sampling interval for distance and time in the dataset.

We found out that the inferred location after swapping was always different from the real one, for all except for the 51 trajectories that did not swapped. However, we inspected some of them and observed that they didn't swapped probably due to the fact that they are outside Beijing, as we can see in Figure 4.

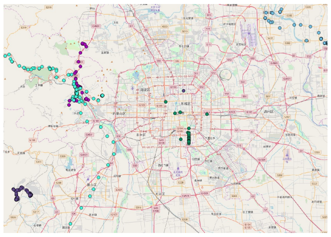


Fig. 4: Some trajectories that did not swapped

4.2 Reidentification by linkage

In this section we simulate an adversary who knows exact locations and timestamps, and tries to reidentify a trajectory in the dataset. We also show that the anonymized trajectories do not always intersect the original ones. This means that, even in the case that the adversary can link the data points that he knows, he may not learn the entire original trajectory.

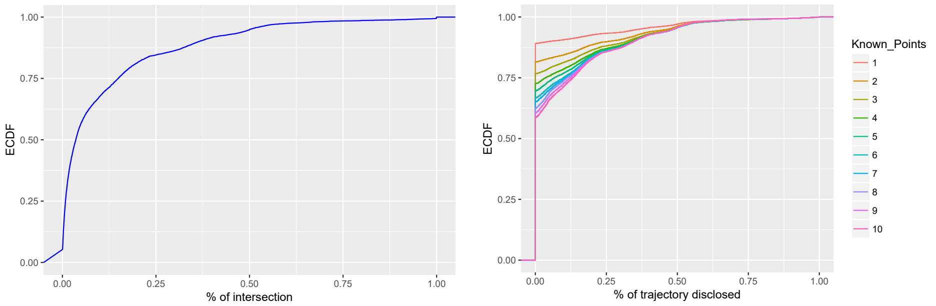
In Figure 5a we represent the empirical cumulative distribution function of the intersection between original and anonymized trajectories, and in Figure 5b the percentage of each trajectory disclosed depending on how many exact spatio-temporal points an adversary knows.

Figure 5a shows that 84% of all the anonymized trajectories intersect in less than 1/4 their corresponding original trajectory, 68% in less than 1/10 and 28% in less than 1/100. Figure 5b shows that adversaries knowing as many as 10 precise spatio-temporal points, are still not able to reidentify 58% of the population, and even when they are able to find the corresponding anonymized trajectory to the one that they are attacking, they will not learn more than 50% of the original trajectory in 95% of the cases. This is in contrast to the results on [17], in which 4 spatio-temporal points are enough to uniquely characterize 95% of the traces, and the traces considered the most difficult to identify can be characterized with 11 points.

5 Conclusions

We have defined and tested a novel algorithm for real-time mobility data anonymization that consists on swapping trajectory segments. In contrast to the *k*-anonymity or differential privacy models for trajectory anonymization, the proposed method does not modify the data, but its association to specific individuals, and it is performed on real time, without the need of having the entire dataset. The proposed protocol tackles both identity and location privacy, and our data model can be adapted to protect either single trajectory positions, as they lose the relation to the individual who has generated the data, or the whole trajectories, since they are mixed among many different peers.

We show that is not possible to infer correctly the home locations after the anonymization and, also, that an adversary who knows exact points of the tra-



(a) intersection between original and anonymized trajectories

(b) attackers capability of inferring trajectories from known locations

Fig. 5: Empirical cumulative distributions

jectory is not able to use them for reidentification, because in most cases they no longer correspond to the anonymized trajectory. And, even in the improbable case that the adversary correctly relates the anonymized trajectory with the original, we have shown that he cannot infer the entire trajectory but just a small part of it.

We have simulated our protocol with an offline algorithm, although, the protocol could be run in real time in which data is transmitted by user devices to our anonymizer that communicates and collaborates with a server. By changing the anonymizer for a group protocol, the protocol could provide security against collusion between the service provider and the anonymizer.

It must be pointed out that swapping cannot be carried out when an individual does not cross anyone in her path. Hence, the proposed technique will not anonymize the individuals who do not cross anyone in their daily activity. However, it is not very common for an individual to spend too much time without meeting someone or going out from home. Moreover, such individuals can be kept outside the database without compromising its utility. The use case considered is for obtaining aggregate mobility data and exact count queries, which neither k-anonymity or differential privacy can provide.

Nevertheless, this comes at the cost of modifying the trajectories and possibly losing individual trajectory mining utility. Future work directions to solve this issue are to add the restriction of non-swapping streets or non-swapping zones for improving the utility to better preserve entire trajectories inside a given street or zone.

Acknowledgments

Julián Salas acknowledges the support of a UOC postdoctoral fellowship. This work is partly funded by the Spanish Government through grant TIN2014-57364-C2-2-R “SMARTGLACIS”, and Swedish VR (project VR 2016-03346).

References

1. Abul, O., Bonchi, F., Nanni, M.: Never walk alone: Uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. pp. 376–385. ICDE '08, IEEE Computer Society, Washington, DC, USA (2008), http://dx.doi.org/10.1109/ICDE.2008.4497446
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. pp. 3–14. ICDE '95, IEEE Computer Society, Washington, DC, USA (1995), http://dl.acm.org/citation.cfm?id=645480.655281
3. Beresford, A.R., Stajano, F.: Location privacy in pervasive computing. IEEE Pervasive Computing 2(1), 46–55 (Jan 2003), http://dx.doi.org/10.1109/MPRV.2003.1186725
4. Beresford, A.R., Stajano, F.: Mix zones: User privacy in location-aware services. In: In Proc. of the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW04). pp. 127–131 (2004)
5. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-time urban monitoring using cell phones: A case study in rome. IEEE Transactions on Intelligent Transportation Systems 12(1), 141–151 (March 2011)
6. Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: De Cristofaro, E., Wright, M. (eds.) Privacy Enhancing Technologies. pp. 82–102. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
7. Chen, R., Fung, B.C., Desai, B.C., Sossou, N.M.: Differentially private transit data publication: A case study on the montreal transportation system. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 213–221. KDD '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2339530.2339564
8. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1082–1090. KDD '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2020408.2020579
9. Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J., Métayer, D.L., Tirtea, R., Schiffner, S.: Privacy and data protection by design - from policy to engineering. Tech. rep., ENISA (2015)
10. Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., Helbing, D.: A planetary nervous system for social mining and collective awareness. The European Physical Journal Special Topics 214(1), 49–75 (Nov 2012), https://doi.org/10.1140/epjst/e2012-01688-9
11. Gidófalvi, G.: Spatio-Temporal Data Mining for Location-Based Services. Ph.D. thesis, Faculties of Engineering, Science and Medicine Aalborg University, Denmark (2007)
12. Golle, P., Partridge, K.: On the anonymity of home/work location pairs. In: Proceedings of the 7th International Conference on Pervasive Computing. pp. 390–397. Pervasive '09, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-01516-8_26
13. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Computing 5(4), 38–46 (Oct 2006)

14. Hoh, B., Gruteser, M.: Protecting location privacy through path confusion. In: Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks. pp. 194–205. SECURECOMM '05, IEEE Computer Society, Washington, DC, USA (2005), http://dx.doi.org/10.1109/SECURECOMM.2005.33
15. Jiang, K., Shao, D., Bressan, S., Kister, T., Tan, K.L.: Publishing trajectories with differential privacy guarantees. In: Proceedings of the 25th International Conference on Scientific and Statistical Database Management. pp. 12:1–12:12. SSDBM, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2484838.2484846
16. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. pp. 193–204. SIGMOD '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1989323.1989345
17. de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific Reports 3 (2013)
18. Pensa, R.G., Monreale, A., Pinelli, F., Pedreschi, D.: Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In: PiLBA (2008), https://air.unimi.it/retrieve/handle/2434/52786/106397/ProceedingsPiLBA08.pdf$#$page=44
19. Reid, D.B.: An algorithm for tracking multiple targets. IEEE Transactions on Automatic Control 24, 843–854 (1979)
20. Romero-Tris, C., Megías, D.: User-centric privacy-preserving collection and analysis of trajectory data. In: Garcia-Alfaro, J., Navarro-Arribas, G., Aldini, A., Martinelli, F., Suri, N. (eds.) Data Privacy Management, and Security Assurance: 10th International Workshop, DPM 2015, and 4th International Workshop, QASA 2015, Vienna, Austria, September 21–22, 2015. Revised Selected Papers, pp. 245–253. Springer International Publishing, Cham (2016), https://doi.org/10.1007/978-3-319-29883-2_17
21. Salas, J., Domingo-Ferrer, J.: Some basics on privacy techniques, anonymization and their big data challenges. Mathematics in Computer Science (2018)
22. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. pp. 188–. PODS '98, ACM, New York, NY, USA (1998), http://doi.acm.org/10.1145/275487.275508
23. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. ICWSM 11, 329–336 (2011)
24. Serjantov, A., Danezis, G.: Towards an information theoretic metric for anonymity. In: Proceedings of the 2nd International Conference on Privacy Enhancing Technologies. pp. 41–53. PET'02, Springer-Verlag, Berlin, Heidelberg (2003), http://dl.acm.org/citation.cfm?id=1765299.1765303
25. Tanimoto, S.L., Itai, A., Rodeh, M.: Some matching problems for bipartite graphs. J. ACM 25(4), 517–525 (Oct 1978), http://doi.acm.org/10.1145/322092.322093
26. Terrovitis, M.: Privacy preservation in the dissemination of location data. SIGKDD Explor. Newsl. 13(1), 6–18 (Aug 2011), http://doi.acm.org/10.1145/2031331.2031334
27. Terrovitis, M., Mamoulis, N.: Privacy preservation in the publication of trajectories. In: Proceedings of the The Ninth International Conference on Mobile Data Management. pp. 65–72. MDM '08, IEEE Computer Society, Washington, DC, USA (2008), https://doi.org/10.1109/MDM.2008.29

28. Xiao, Y., Xiong, L.: Protecting locations with differential privacy under temporal correlations. In: Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security. pp. 1298–1309. CCS '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2810103.2813640
29. Yuan, J., Zheng, Y., Xie, X., Sun, G.: Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 316–324. KDD '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2020408.2020462
30. Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y.: T-drive: Driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 99–108. GIS '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1869790.1869807
31. Zang, H., Bolot, J.: Anonymization of location data does not work: A large-scale measurement study. In: Proceedings of the 17th Annual International Conference on Mobile Computing and Networking. pp. 145–156. MobiCom '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2030613.2030630
32. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th International Conference on World Wide Web. pp. 791–800. WWW '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1526709.1526816
33. Zhou, B., Pei, J., Luk, W.: A brief survey on anonymization techniques for privacy preserving publishing of social network data. SIGKDD Explor. Newsl. 10(2), 12–22 (Dec 2008), http://doi.acm.org/10.1145/1540276.1540279