Japanese Journal of Statistics and Data Science

Swapping trajectories with a sufficient sanitizer

--Manuscript Draft--

Manuscript Number:		
Full Title:	Swapping trajectories with a sufficient sanitizer	
Article Type:	Original Paper	
Funding Information:	Royal Swedish Academy of Sciences (VR 2016-03346)	Dr Vicenç Torra
	Blavatnik Interdisciplinary Cyber Research Center (IL) (Privacy-Utility in Mobility)	Dr. Raazesh Sainudiin
	Spanish Government Grant "SMARTGLACIS" (TIN2014-57364-C2-2-R)	Dr Julián Salas
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, such as differential privacy, are based on adding noise in a way that the presence of an individual cannot be inferred from the data. In this paper, we consider a different anonymization approach by swapping partial trajectories, formalize the concept of sufficient sanitizer and show that the proposed method (SwapMob), is a sufficient sanitizer for various statistical decision problems. That is, it preserves the aggregate information of the spatial database in the form of sufficient statistics and also provides privacy to the individuals. We measure the privacy provided by SwapMob as the Adversary Information Gain, which measures the capability of an adversary to leverage his knowledge of a given data point to infer a larger segment of the sanitized trajectory. We test the utility of the data obtained after applying SwapMob sanitization in terms of Origin-Destination matrices, a fundamental tool in transportation modelling.	
Corresponding Author:	Raazesh Sainudiin Uppsala Universitet Matematiska institutionen Uppsala, SWEDEN	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Uppsala Universitet Matematiska institutionen	
Corresponding Author's Secondary Institution:		
First Author:	Julián Salas	
First Author Secondary Information:		
Order of Authors:	Julián Salas	
	David Megίas	
	Vicenç Torra	
	Marina Toger	
	Joel Dahne	
	Raazesh Sainudiin	
Order of Authors Secondary Information:		

Author Comments:	
Additional Information:	
Question	**Response**
Please verify that you have included a conflict of interest statement in your manuscript. A conflict of interest exists whenever an author has a financial or personal relationship with a third party whose interests could be positively or negatively influenced by the article's content. This should be added in a separate section before the reference list. If no conflict of interest exists for all participating authors, the corresponding author should use the following wording: On behalf of all authors, the corresponding author states that there is no conflict of interest.	
Does this manuscript belong to a special feature?	Yes
Please select the special feature your manuscript belongs to. as follow-up to "Does this manuscript belong to a special feature?"	Information Theory and Statistics

Noname manuscript No.
(will be inserted by the editor)

Swapping trajectories with a sufficient sanitizer

Julián Salas · David Megías · Vicenç Torra · Marina Toger · Joel Dahne · Raazesh Sainudiin

Received: date / Accepted: date

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, such as differential privacy, are based on adding noise in a way that the presence of an individual cannot be inferred from the data.

Julián Salas
Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Barcelona, Spain.
CYBERCAT-Center for Cybersecurity Research of Catalonia.
E-mail: jsalaspi@uoc.edu

David Megías
Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Barcelona, Spain.
CYBERCAT-Center for Cybersecurity Research of Catalonia.
E-mail: dmegias@uoc.edu

Vicenç Torra
Hamilton Institute, Maynooth University, Maynooth, Ireland.
School of Informatics, University of Skövde, Skövde, Sweden.
E-mail: vtorra@his.se

Marina Toger
Department of Economic and Cultural Geography, Uppsala University,
E-mail: marina.toger@kultgeog.uu.se

Joel Dahne
Department of Mathematics, Uppsala University, Uppsala, Sweden.
E-mail: joel@dahne.eu

Raazesh Sainudiin
Department of Mathematics, Uppsala University, Uppsala, Sweden.
E-mail: raazesh.sainudiin@math.uu.se

In this paper, we consider a different anonymization approach by swapping partial trajectories, formalize the concept of sufficient sanitizer and show that the proposed method (SwapMob), is a sufficient sanitizer for various statistical decision problems. That is, it preserves the aggregate information of the spatial database in the form of sufficient statistics and also provides privacy to the individuals. We measure the privacy provided by SwapMob as the Adversary Information Gain, which measures the capability of an adversary to leverage his knowledge of a given data point to infer a larger segment of the sanitized trajectory. We test the utility of the data obtained after applying SwapMob sanitization in terms of Origin-Destination matrices, a fundamental tool in transportation modelling.

Keywords Privacy Preserving Trajectory Mining · Origin-Destination matrices · Trajectory anonymization · Intelligent Transportation Systems · Sufficient Sanitizer · Mobility Data Mining

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 (Terrovitis, 2011, for e.g.).

Moreover, as mentioned by Giannotti et al. (2012), 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. Terrovitis and Mamoulis (2008) consider a discrete spatial domain, say for e.g., spatial information is given in terms of addresses in a city map. Hence, the user trajectories are expressed as sequences of points of interest (POIs). They present the use case of the RFID cards from the Octopus[1] 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 is obtained in Pensa et al. (2008), transforming sequences by adding, deleting, or substituting some points of the trajectory, while preserving also frequent sequential patterns (Agrawal and Srikant, 1995) obtained by mining the anonymized data.

Hoh and Gruteser (2005) and Hoh et al. (2006) 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 (Reid, 1979) subsequent location samples may be linked to an individual that is periodically reporting his anonymized location information.

Hoh et al. (2006) 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.

To prevent adversaries from tracking complete individual paths, Hoh and Gruteser (2005) 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 the concept of Mix Zones introduced in Beresford and Stajano (2003). 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 disrupt the linkage of the incoming and outgoing path segments to the same specific user.

[1] 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 mix zone.

To measure location privacy, Beresford and Stajano (2004) define the anonymity set as the group of people visiting the mix zone during the same time interval. 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 (2003) proposed for anonymous communications which considers the varying probabilities of users sending and receiving messages through a network of mix nodes.

This is modeled by Beresford and Stajano (2004) as a movement matrix which represents 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 can be used to find the most probable mapping among incoming and outgoing pseudonyms. However, since the solution to many restricted matching problems including this one is NP-hard (Tanimoto et al., 1978), Beresford and Stajano (2004) describe a method for achieving partial solutions.

Other approaches to provide privacy to LBS include the posibility of individual specification of privacy preferences (Damiani et al., 2009), some are based on Private Information Retrieval (PIR) and Oblivious Transfer (Paulet et al., 2014), or are a combination of cloaking regions and PIR (Ghinita et al., 2011), however they use cryptographic primitives and are focused on preventing the disclosure of points of interest.

An approach that does not consider middleware to obtain location privacy is proposed by Gidófalvi (2007, Chapter 9). 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, in the final step (the data reporting stage) clients send anonymous partial trajectories to the server, it filters all the synthetic trajectory data generated during the process and recovers the original trajectory.

One of the advantages of performing trajectory anonymization on the user side, as in Romero-Tris and Megías (2016) and Romero-Tris and Megías (2018), is that the anonymization process is no longer centralized. Thus data subjects gain control, transparency and more security for their data. They leverage the concept of k-anonymity for trajectories, similarly to Abul et al. (2008), that propose the (k, δ)-anonymity model, which consists of 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 (Samarati and Sweeney, 1998) and it may be related to k-anonymity for dynamic databases (Salas and Torra, 2018) if we consider that the records of the dynamic database represent locations. Also the concept of differential privacy (Dwork et al., 2006) has been extended from databases to many other types of data. For a brief overview of privacy protection techniques and a discussion of k-anonymity and differential privacy models in different frameworks cf. Salas and Domingo-Ferrer (2018).

Chen et al. (2012) consider a differential privacy model for transit data publication using data from the Société de Transport de Montréal (STM). The data are modeled sequentially in a prefix tree that represents all the sequences by grouping those 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 $\tilde{D}$ 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 (Agrawal and Srikant, 1995).

Jiang et al. (2013) present another ϵ-differentially private mechanism for publishing trajectories called SDD (Sampling Distance and Direction). They focus on ship trajectories with known starting and terminal point, with same number of points, and consider differential privacy when two trajectories differ at exactly one location.

Xiao and Xiong (2015) propose a differentially private algorithm for location privacy that follows a discussion on the different notions of adjacency used for differential privacy (Chatzikokolakis et al., 2013; Kifer and Machanavajjhala, 2011, for e.g.). 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 a sanitization 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. A preliminary version of this study appeared in Salas et al. (2018). The main additions to this study are the formalization of sufficent sanitizer for various models, sanitizers that preserve Origin-Destination matrices, an evaluation of the trade off between privacy and utility. We also consider a new measure of Information Gain for an attacker that is different from the measures in previous version.

The rest of the paper is organized as follows. In Section 3 we formalize the concept of sufficient sanitizer and provide some examples of sufficient statistcs related to traffic engineering and transportation planning. In Section 4 we define our algorithm and discuss some of its properties related to privacy and utility. In Section 5, we evaluate our method also when including the additional guarantee of preserving Origin-Destination matrices. We finish with some conclusions and future work on Section 6.

3 Sufficient Sanitizer

Recall that the sufficient statistics T of data X under a statistical experiment, i.e., a family of probability models $\{P_\theta : \theta \in \mathbf{\Theta}\}$, that is parametrized by $\theta \in \mathbf{\Theta}$, contains all the information in the data about θ. More formally, $T(X) = t$ is a sufficient statistic for the underlying parameter θ if the conditional probability $P_\theta(X|T(X) = t)$ is independent of θ. Intuitively speaking, all the information in the data about the typically unknown parameter of the probability model is captured by the sufficient statistic. Therefore, sanitizing 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 $\tilde{\mathcal{D}}$, i.e., $\mathcal{S}(d) : \mathbb{D} \to \tilde{\mathcal{D}}$, in such a way that the data in $\tilde{\mathcal{D}}$ has additional privacy guarantees than in $\mathbb{D}$. Thus, this notion englobes the notion of k-anonymity, differential privacy, and many other privacy enhancing technologies. A sanitizer may or may not preserve the sufficient statistics in the data for a given statistical experiment. A *sufficient sanitizer* preserves the sufficient statistics.

Next, we give three concrete examples of sufficient statistics that have utility in decision problems routinely faced in traffic engineering, city and transportation planning, etc. We end this section with a discussion on General Markov models and sufficient sanitizers.

3.0.1 State Counts:

One of the simplest statistical experiments for mobility data can be based on an independent and identical distribution for the probability of being found in location or state i from among $k + 1$ states based on a labelled partition of the support set of the trajectories into $k + 1$ cells or states given by $[k] := \{0, 1, \ldots, k\}$. For such a simple experiment $\{P_\theta : \theta \in \triangle^k\}$, i.e., P_θ is a discrete probability distribution specified by the parameter θ taking values in $\triangle^k := \{\theta \in \mathbb{R}^{k+1} : \sum_{i=0}^{k} \theta_i = 1, \theta_i \geq 0, i \in [k]\}$, the probability k-simplex. A consistent nonparametric estimate of θ is obtained from the relative frequency

Table 1: An Origin Destination Matrix from a spatial interaction survey

		Destinations j			
		Uppsala	Stockholm	Arlanda	Departures
Origins i	Uppsala	2000	5	20	2025
	Stockholm	10	100	10	120
	Arlanda	20	5	0	25
	Arrivals	2030	110	30	2170

of visits to each state in $[k]$ and its sufficient statistic is simply $\{N_i : i \in [k]\}$, where N_i is merely the frequency or count of the number of visits to state i.

3.0.2 State Transition Counts:

A more useful statistical experiment for trajectory data is the time-homogeneous Markov chain model of independent random transitions. This model is more general the previous one, since it allows the probability of the next location to depend on that of the current location. Here $\{P_\theta : \theta \in (\triangle^k)^k\}$, i.e., P_θ is the transition probability matrix of a Markov chain based on a partition of the support set of the trajectories into $k+1$ labelled cells or states given by $[k]$. Recall that the transition counts $N_{i,j}$ between states i and j for each pair $(i, j) \in [k]^2$ is the sufficient statistics for such a simple Markov chain model, as it will allow us to nonparametrically estimate the transition matrix itself.

3.0.3 Origin-Destination Matrices:

Origin-Destination Matrices (ODMs) are routinely used in transportation modelling to depict travel demand. Traffic flows can be estimated as part of trip generation modelling using Origin-Destination (OD) demand matrix, infrastructure network capacity and traffic controls. OD trip generation models serve as basis for transport planning, construction, performance assessment, and as such have potential to affect regional economies.

Although ODMs can be more general, we consider an ODM based on n states that can be the origin i and/or the destination j. Such an ODM shown in Table 1 is a matrix of size $(n+1) \times (n+1)$ containing flow values N_{ij}, such as number or share of trips from i to j (Rodrigue et al., 2009). The last row contains total arrivals to each destination j from all origins, the last column contains the total departures from each origin i to all destinations, and the bottom right element contains the total flows in the model (Evans, 1970).

ODM are constructed based on estimations from travel studies as part of traffic census: field, online and telephone traffic surveys, traffic volume counts (Robillard, 1975), check-point intercept interviews, license plate and other video analyses, *etc.* Automatically generated data (Iqbal et al., 2014, e.g. CDR) are increasingly used as a base for constructing ODMs, reducing survey costs and improving accuracy of route choice estimations. Thus, a sufficient sanitizer for ODMs from trajectory data (as SwapMob), can allow for the utility to be gained from ODM while preserving the privacy of the individual associated with the trajectory.

ODM parameters include: cut-off departure time from Origins, cut-off arrival time to Destinations, mode of transportation, and spatial resolution or aggregation level for Origins and Destinations. In Section 5.3, we empirically assess the loss of privacy under a given metric as this spatial resolution varies for a single ODM.

3.0.4 General Markov models:

More sophisticated Markov chain models, including those that allow dependence on past few states or those that allow the transition probabilities to depend on time with more involved sufficient statistics, can in principle be treated with the basic ideas illustrated here using simple but useful Markov chain models of mobility. Thus, any subsequent decision problem (eg. traffic flow prediction from mobility simulations based on the learnt Markov chain model), based on *sufficient sanitizers* that preserve the sufficient statistic for the model can allow for optimal decisions under the model for a desired level of privacy.

4 Proposed method: SwapMob

We propose a method for sanitization 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 4.1.1. 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.

4.1 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.

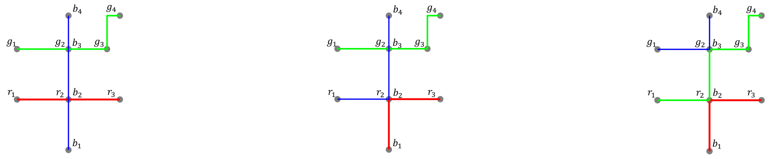
Note that, in case that the number of elements m is odd, to generate a matching we must leave out one element and that all possible random matchings can be generated following our procedure.

4.1.1 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 also be done as a permutation akin to Beresford and Stajano (2004). 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 $Sw(T_g) = (r_1, r_2, b_3, g_3, g_4)$.



(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 $Sw(T_i)$
```

4.2 SwapMob sanitizer

We follow an architecture similar to Hoh et al. (2006) in which a Trusted Third Party (TTP) knows the vehicles' identities but cannot 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 sanitizer) 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

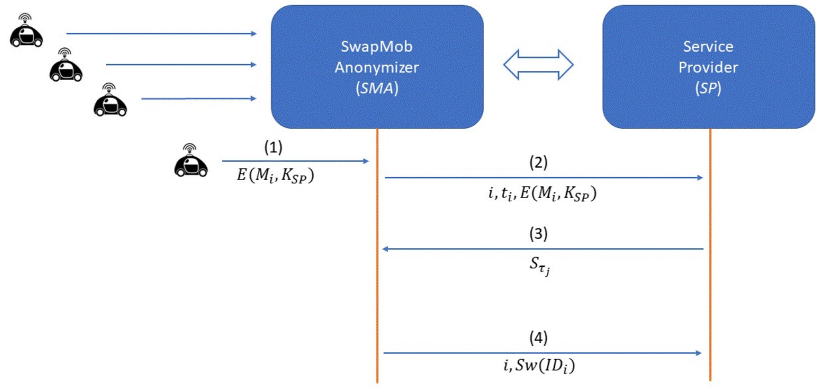


Fig. 2: Architecture of our system

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 pre-defined threshold χ during the interval of time τ_j back to SwapMob, more formally $S_{\tau_j} = \{i, i' : d(l_i, l_{i'}) < \chi$ 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.

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 sanitized 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 swaps 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 sanitized by SwapMob (Step 4).

Our system, can be applied for the use case proposed in Beresford and Stajano (2003), 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 Beresford and Stajano (2003) 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.

4.3 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 their data to their real identity.

Other notions of privacy are defined depending on the context, which may be of statistical databases (Danezis et al., 2015), networks (Zhou et al., 2008), or geo-located data.

By identifying the POIs of an individual, it is possible to infer their 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 their schedule, predict their future locations, learn their past locations and possibly even infer their personal relations by observing frequent or periodic co-locations. Moreover, such habits and locations can be easily used to reidentify the individuals behind the data, as it has been shown on previous studies on anonymity of home/work location.

Regarding this topic, Golle and Partridge (2009) showed that the sizes of the anonymity set were 1, 21 or 34980 for 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), respectively. 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. Zang and Bolot (2011) 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, it 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 de Montjoye et al. (2013), 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 privacy by dissociating the segments of trajectories from the subject that generated them.

4.4 Utility of swapped data as privacy-preserving decisions

In this paper we are assuming that the interest of using data sanitized by SwapMob is for making mobility maps and predictions that may be useful for intelligent transportation systems and for planning in a city. Our main contribution here is to formalize the relationship between statistical decision procedures that are used to extract utility from the data on one hand and sanitizers that attempt to guarantee a particular measure of privacy via sanitization on the other.

4.4.1 Current examples:

As Hoh and Gruteser (2005) proposed, 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 Calabrese et al. (2011), 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 7,268 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 sensors that could be carried by individuals, sanitized and transmitted to a service provider via SwapMob.

Another example is the offline mining in Yuan et al. (2011) representing the knowledge from taxi-drivers as a landmark graph could be done with SwapMob sanitized 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.

4.4.2 SwapMob is a Sufficient Sanitizer:

In general, lossless maps of flows, up to a statistical experiment and its sufficient statistics, encoded by *sufficiently sanitized trajectories* can be obtained by using SwapMob at several aggregation levels and time resolutions specified by χ and τ, respectively. Next we show that unlike k-anonymity and ϵ-differential privacy based approaches, SwapMob does indeed preserve the sufficient statistics of counts, transition counts and Origin-Destination matrix (ODM) – the three sufficient statistics with their corresponding statistical experiments described in Section 3.

It is easy to see that the SwapMob sanitizer for a given time interval τ and spatial resolution specified by χ, preserves the sufficient statistics of counts in each cell or state given by the χ-specified spatial partition. First, the swapping operation within each cell only swaps random pairs of trajectories within it and thus leaves the counts invariant. Also, the points of entry and exit for each trajectory for a given spatio-temporal cell is preserved as the swapping operation only happens across random pairs of trajectories inside the cell. Thus, the number of transitions between any two spatio-temporal cells will also be preserved. This actually preserves the sufficient statistics for the time-inhomogeneous Markov chain and not merely that of the time-homogeneous Markov chain. Note that, although we have talked about discrete time Markov chains specified by units of τ, we can just as easily generalize the underlying models to continuous-time Markov chains by appropriate projections and the use of timestamp information in the trajectories.

4.5 Privacy measure: Adversary Information Gain

The level of privacy provided by an algorithm is commonly measured with the a priori guarantees of k-anonymity or ϵ-differential privacy, in which tuning the parameters of k and ϵ increases or decreases the level of privacy. Also, the concept of unicity from de Montjoye et al. (2013) may be related to privacy of mobility data, because an adversary having auxiliar (external) information may uniquely identify an individual in the database by relating such information with a unique record.

Our assumption here, is that the adversary knows a precise point (location and timestamp) that uniquely identifies an individual. This means that the adversary can link such point to the sanitized individuals' data.

This is in contrast to the three methods for measuring privacy that we have just mentioned. Firstly, in k-anonymity there are not unique Quasi-Identifiers and their combinations are repeated at least k-times. Secondly, for ϵ-differential privacy, the presence of an individual in the database is not revealed (up to ϵ). Finally, in the unicity tests from de Montjoye et al. (2013) the location and timestamp data is published up to some resolution.

Therefore, we define the *Adversary Information Gain (AIG)*, as the fraction of the original trajectory that is disclosed by an adversary who knows a point of it. Or equivalently, the amount of additional information that an adversary can gain after linking a known point to a trajectory in the sanitized database. Here, we consider Adversary Information Gain in terms of the number of measurements, but elapsed time and distance traveled would both be other natural choices.

We consider that an adversary can propagate his knowledge of one data point to the whole segment in-between swaps: we assume that the adversary follows the trajectory from the known measurement forward in time until the next swap occurs and also backwards in time until the previous swap has occurred. The adversary will know with certainty that such segment belongs to the same individual, since no swaps were performed on it, but after the swaps will not know which path belongs to the original trajectory. We will show, in Section 5.2 that in most cases, only a small fraction of the full trajectory is disclosed.

5 Empirical evaluation

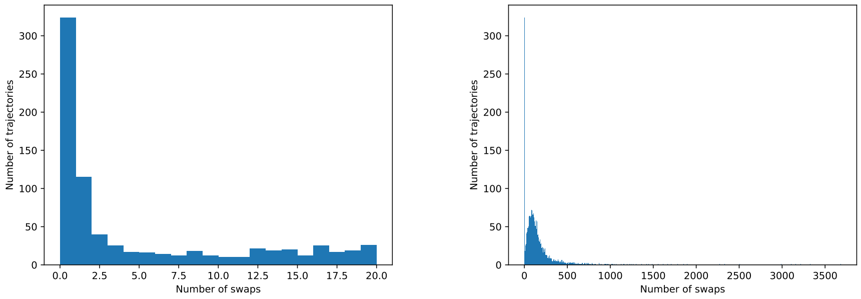
We tested our algorithm on the T-drive dataset (Yuan et al., 2010, 2011) which contains the GPS trajectories of 10,357 taxis during the period of Feb. 2 to Feb. 8, 2008 within Beijing. The total number of points in this dataset is about 15 million and the total distance of the trajectories reaches nearly 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 measurement contains the following data: taxi id, date time, longitude, latitude.

5.1 Applying SwapMob on the data

Before applying SwapMob to the dataset we perform some cleaning of the data. We begin by removing all measurements for which the latitude and longitude is outside the box $[115, 117] \times [39, 41]$, these measurements are far outside Beijing and most of them have both latitude and longitude equal to zero, which indicates that the measurement is most likely not valid. We then remove all trajectories which have no measurements belonging to them at all. This removes 756,561 invalid measurements and 77 trajectories that have no measurements in this area, and we end up with 10280 trajectories and 16,906,423 measurements.

For applying SwapMob we consider two taxis co-located if they are in the same spatio-temporal cell – the spatial cell is given by a square of side-length 0.001 degrees ($\chi = 0.001$), about 111 meters, and the temporal interval is specified by being in the same minute ($\tau = 60$). Note that this is about 6 and 3 times less than the average sampling interval for distance and time, respectively, in the dataset.

With this we get that the number of possible swaps between all the trajectories is 641,262. The average number of swaps for each trajectory is 137.



(a) Frequency histogram of number of swaps for all the 772 trajectories participating in less than 20 swaps.

(b) Frequency histogram of number of swaps for all trajectories.

Fig. 3: Frequency histograms of number of swaps for trajectories.

Of all the 10,280 trajectories 9508, 92%, take part in at least 20 swaps, for the 772 trajectories with less than 20 swaps we have the distribution seen in Figure 3a. We see that 324 trajectories do not participate in any swaps at all, however 265 of these trajectories have less than 10 measurements (compared to an average of more than 1000). In Figure 3b we can also see the distribution for all trajectories, here we can see that most trajectories have less than 500 swaps but a few participate in thousands of swaps.

We can conclude that most of the trajectories participate in several swaps, only a small amount participate in less than 20 swaps and that most of the trajectories that don't participate in any swap have very few measurements.

5.2 Adversary Information Gain

As we explained in Section 4.5, to measure Adversary Information Gain (AIG), we have to measure the length of the segments between the swaps. Since every trajectory participates in an average of 137 swaps and is thus split into an average of 138 segments. Knowing one point the adversary will learn one of these segments. For each trajectory we can look at the longest such segment and compare its length to that of the whole trajectory. We plotted the distribution function of such measure in Figure 4.

Figure 4 shows that for more than 75% of all trajectories, the AIG is less than 0.2, for 90% of the trajectories it is less than 0.4. For most trajectories, an adversary will thus learn only a small fraction of the original trajectory.

We can also consider the case when the adversary knows several points of the trajectory. If the extra points lie on the same segment of the trajectory as the first one, no more knowledge is gained. However, if the extra points lie on other segments of the trajectory, the adversary learns also these and more of the trajectory is disclosed. If the adversary knows several segments of the

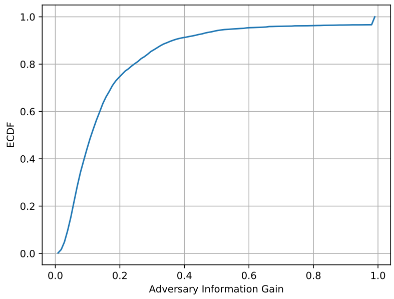


Fig. 4: Cummulative Distribution Function of Adversary Information Gain.

trajectory, it can also try to infer which way was taken between them, and thus learning more than just the disclosed segments. Analysing this is however outside the scope of this article and left as further research.

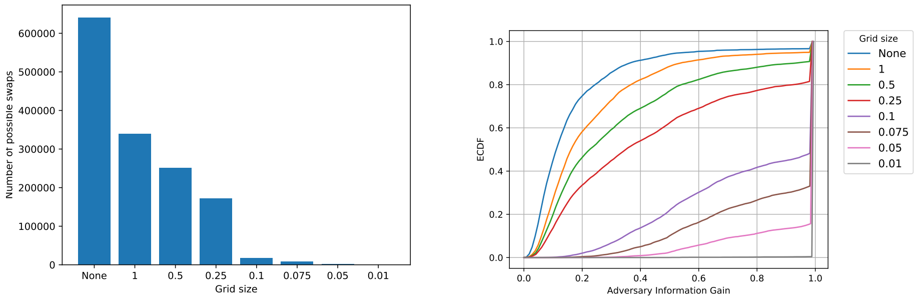
5.3 Privacy when preserving the Origin-Destination Matrix

We now consider the effects on the privacy measures of restricting swapping to preserve the Origin-Destination Matrix (ODM), introduced in Section 3. That is, for two trajectories to swap they have to share the same starting location or origin and ending location or destination (up to some scale). This is in addition to the earlier requirements for allowing a swap.

We define the states or locations used in the ODM by the labelled cells or states in a grid obtained by partitioning the city into equally sized squares (in units of degrees). The resolution of the grid is given by the width or side-length of the squares in degrees of latitude and longitude (1 degree of latitude is about 111,000 m in Beijing), a lower width then corresponds to a finer grid. The start location or origin for a trajectory will be given by the square that its first measurement belongs to and the end location or destination by the square that its last measurement belongs to. In general, it might be more appropriate to have start and end locations to be determined by the location of the trajectory at a certain time of the day. However, for keeping it simple, we choose to use only the first and last measurement. More generally, our approach allows for an arbitrary set of subsets of the city and arbitrary time-intervals to specify Origins and Destinations in a single ODM, or even a sequence of ODMs, but we use a simple grid-based partition at different spatial resolutions to illustrate the effects on privacy here.

We will analyze how the preserved privacy changes when we go from having no grid to having a very coarse grid and then making it finer and finer.

In Figure 5a we see how the number of possible swaps change when we make the grid finer. At first, we have the number of swaps without a grid (same as Figure 4) and then we have the numbers for grids of squares of the



(a) The number of possible swaps depending on the grid size. Going from having no grid to a very fine grid.

(b) Adversary Information Gain for different grid sizes.

Fig. 5: Preserved privacy when also preserving the ODM

given height. The largest height used is 1 degree, about 11100 meters, which can be compared to the distance of 0.001 degrees, about 111 meters, that we used for determining if two trajectories are swappable, i.e. close enough to be swapped. This grid splits the city into four parts, of which two contain most of the measurements. On the other extreme, the finest grid is made up of squares of width 0.01 degrees or about 1110 meters, and this is 10 times the proximity threshold for swappability. We see that the number of possible swaps quickly decreases as the grid gets finer. Even at the coarsest grid, the number of swaps is halved compared to having no grid. When the grid height goes below 0.1 we have very few possible swaps.

Since the preserved privacy heavily depends on the number of possible swaps, we expect it to quickly decrease as the grid becomes finer. We can confirm this by computing the same measurement as in Section 5.2 for each grid size, see Figure 5b. The line corresponding to having no grid is the same as in Figure 4 and we can see that the privacy decreases as the grid becomes finer.

We can conclude that the preserved privacy is greatly influenced by the grid size of the ODM. If the grid is too fine almost no swaps occur and the SwapMob algorithm is not efficient. For very coarse grids the privacy is still reduced but depending on the application it could still be considered acceptable. Furthermore, by defining a sequence of ODMs, say $(M_1, M_2, \ldots, M_m)$, specified by arbitrary subsets of space and intervals of time $[\underline{t}_{i,o}, \overline{t}_{i,o}]$ and $[\underline{t}_{i,d}, \overline{t}_{i,d}]$ for each M_i, one can increase privacy by increasing the number of swappable trajectories. Such a sequence of ODMs should generally be of greater utility for certain decision problems. We defer a thorough investigation of sufficient sanitizers that preserve sufficient statistics for such sequences of ODMs across spatial and temporal resolutions in a principled manner for future research.

6 Conclusions

We have defined and tested a novel algorithm for mobility data sanitization that consists of swapping trajectory segments. In contrast to the k-anonymity or differential privacy models for trajectory sanitization, the proposed method does not modify the data, but its association to specific individuals, and it is performed in 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 for the T-drive dataset a high number of swaps occur and most trajectories participate in several swaps. We can see that the original trajectories are split up into many smaller segments and that it makes it hard even for an adversary who knows exact points of the trajectory to infer much of the full trajectory.

This is expressed as the Adversary Information Gain, which we formally defined to measure the capability of an adversary to leverage his knowledge of a given exact point, to infer a larger segment of the sanitized trajectory.

However, when adding the restriction of preserving the ODM, the number of swaps quickly decreases as the grid is made finer and it becomes much easier for an adversary to infer a big portion of the trajectory of an individual knowing only a few measurements. This is the natural tradeoff between the *societal utility gained* through the preservation of the ODM in decision problems, where ODM is a sufficient statistic, and the *individual privacy lost* by the sufficient sanitizer.

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 sanitizer that communicates and collaborates with a server. By changing the sanitizer for a group protocol, the protocol could provide security against collusion between the service provider and the sanitizer.

It must be mentioned that swapping cannot be carried out when an individual does not come close to anyone along their path. Hence, the proposed technique will not protect an individual who does not proximally encounter anyone in their daily activity. We consider that 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, e.g., in the case of T-drive data 265 of these 324 trajectories have less than 10 measurements (compared to an average of more than 1000).

The use case considered is for obtaining aggregate mobility data that preserve sufficient statistics for various statistical decision problems used in traffic engineering and city planning, including exact count queries, transition count queries and ODM queries, which neither k-anonymity nor differential privacy cannot formally guarantee.

As we have shown, there is a clear tradeoff between preserving a particular statistic, such as ODM, and the privacy provided by this algorithm. While swapping without constraints provides the highest level of privacy, it comes at the cost of 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 to better preserve entire trajectories inside a given street or zone. A formal privacy-preserving decision-theoretic framework based on probabilistic models and statistical experiments for co-trajectories that can be integrated across multiple spatial and temporal resolutions needs further investigations, especially when the computational setting becomes distributed to handle mobility data at a massive scale.

Acknowledgements 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", Swedish VR (project VR 2016-03346) and a grant to Raazesh Sainudiin from Blavatnik Interdisciplinary Cyber Research Center.

References

Abul O, Bonchi F, Nanni M (2008) Never walk alone: Uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, IEEE Computer Society, Washington, DC, USA, ICDE '08, pp 376–385, DOI 10.1109/ICDE.2008.4497446, URL http://dx.doi.org/10.1109/ICDE.2008.4497446

Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, IEEE Computer Society, Washington, DC, USA, ICDE '95, pp 3–14, URL http://dl.acm.org/citation.cfm?id=645480.655281

Beresford AR, Stajano F (2003) Location privacy in pervasive computing. IEEE Pervasive Computing 2(1):46–55, DOI 10.1109/MPRV.2003.1186725, URL http://dx.doi.org/10.1109/MPRV.2003.1186725

Beresford AR, Stajano F (2004) 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

Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: A case study in rome. IEEE Transactions on Intelligent Transportation Systems 12(1):141–151, DOI 10.1109/TITS.2010.2074196

Chatzikokolakis K, Andrés ME, Bordenabe NE, Palamidessi C (2013) Broadening the scope of differential privacy using metrics. In: De Cristofaro E, Wright M (eds) Privacy Enhancing Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 82–102

Chen R, Fung BC, Desai BC, Sossou NM (2012) 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, ACM, New York, NY, USA, KDD ’12, pp 213–221, DOI 10.1145/2339530.2339564, URL http://doi.acm.org/10.1145/2339530.2339564

Damiani ML, Bertino E, Silvestri C (2009) Protecting location privacy against spatial inferences: The probe approach. In: Proceedings of the 2Nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS, ACM, New York, NY, USA, SPRINGL ’09, pp 32–41, DOI 10.1145/1667502.1667511, URL http://doi.acm.org/10.1145/1667502.1667511

Danezis G, Domingo-Ferrer J, Hansen M, Hoepman JH, Métayer DL, Tirtea R, Schiffner S (2015) Privacy and data protection by design - from policy to engineering. Tech. rep., ENISA

Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T (eds) Theory of Cryptography, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 265–284

Evans A (1970) Some properties of trip distribution methods. Transportation Research 4(1):19 – 36, DOI https://doi.org/10.1016/0041-1647(70)90072-9, URL http://www.sciencedirect.com/science/article/pii/0041164770900729

Ghinita G, Kalnis P, Kantarcioglu M, Bertino E (2011) Approximate and exact hybrid algorithms for private nearest-neighbor queries with database protection. GeoInformatica 15(4):699–726, DOI 10.1007/s10707-010-0121-4, URL https://doi.org/10.1007/s10707-010-0121-4

Giannotti F, Pedreschi D, Pentland A, Lukowicz P, Kossmann D, Crowley J, Helbing D (2012) A planetary nervous system for social mining and collective awareness. The European Physical Journal Special Topics 214(1):49–75, DOI 10.1140/epjst/e2012-01688-9, URL https://doi.org/10.1140/epjst/e2012-01688-9

Gidófalvi G (2007) Spatio-temporal data mining for location-based services. PhD thesis, Faculties of Engineering, Science and Medicine Aalborg University, Denmark

Golle P, Partridge K (2009) On the anonymity of home/work location pairs. In: Proceedings of the 7th International Conference on Pervasive Computing, Springer-Verlag, Berlin, Heidelberg, Pervasive ’09, pp 390–397, DOI 10.1007/978-3-642-01516-8_26, URL http://dx.doi.org/10.1007/978-3-642-01516-8_26

Hoh B, Gruteser M (2005) Protecting location privacy through path confusion. In: Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks, IEEE Computer Society, Washington, DC, USA, SECURECOMM ’05, pp 194–205, DOI 10.1109/SECURECOMM.2005.33, URL http://dx.doi.org/10.1109/SECURECOMM.2005.33

Hoh B, Gruteser M, Xiong H, Alrabady A (2006) Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Computing 5(4):38–46, DOI 10.1109/MPRV.2006.69

Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin–destination matrices using mobile phone call data. Transportation

Research Part C: Emerging Technologies 40:63–74

Jiang K, Shao D, Bressan S, Kister T, Tan KL (2013) Publishing trajectories with differential privacy guarantees. In: Proceedings of the 25th International Conference on Scientific and Statistical Database Management, ACM, New York, NY, USA, SSDBM, pp 12:1–12:12, DOI 10.1145/2484838.2484846, URL http://doi.acm.org/10.1145/2484838.2484846

Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '11, pp 193–204, DOI 10.1145/1989323.1989345, URL http://doi.acm.org/10.1145/1989323.1989345

de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The privacy bounds of human mobility. Scientific Reports 3

Paulet R, Kaosar MG, Yi X, Bertino E (2014) Privacy-preserving and content-protecting location based queries. IEEE Trans on Knowl and Data Eng 26(5):1200–1210, DOI 10.1109/TKDE.2013.87, URL https://doi.org/10.1109/TKDE.2013.87

Pensa RG, Monreale A, Pinelli F, Pedreschi D (2008) Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In: PiLBA, URL https://air.unimi.it/retrieve/handle/2434/52786/106397/ProceedingsPiLBA08.pdf$#$page=44

Reid DB (1979) An algorithm for tracking multiple targets. IEEE Transactions on Automatic Control 24:843–854

Robillard P (1975) Estimating the o-d matrix from observed link volumes. Transportation Research 9(2):123–128, DOI 10.1016/0041-1647(75)90049-0, URL http://www.sciencedirect.com/science/article/pii/0041164775900490

Rodrigue JP, Comtois C, Slack B (2009) The geography of transport systems. Routledge, URL http://transportgeography.org/

Romero-Tris C, Megías D (2016) 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, Springer International Publishing, Cham, pp 245–253, DOI 10.1007/978-3-319-29883-2_17, URL https://doi.org/10.1007/978-3-319-29883-2_17

Romero-Tris C, Megías D (2018) Protecting privacy in trajectories with a user-centric approach. ACM Trans Knowl Discov Data 12(6):67:1–67:27, DOI 10.1145/3233185, URL http://doi.acm.org/10.1145/3233185

Salas J, Domingo-Ferrer J (2018) Some basics on privacy techniques, anonymization and their big data challenges. Mathematics in Computer Science 12(3):263–274, DOI 10.1007/s11786-018-0344-6, URL https://doi.org/10.1007/s11786-018-0344-6

Salas J, Torra V (2018) A general algorithm for k-anonymity on dynamic databases. In: Garcia-Alfaro J, Herrera-Joancomartí J, Livraga G, Rios R (eds) Data Privacy Management, Cryptocurrencies and Blockchain Technology, Springer International Publishing, Cham, pp 407–414

Salas J, Megías D, Torra V (2018) Swapmob: Swapping trajectories for mobility anonymization. In: Domingo-Ferrer J, Montes F (eds) Privacy in Statistical Databases, Springer International Publishing, Cham, pp 331–346

Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM, New York, NY, USA, PODS '98, pp 188–, DOI 10.1145/275487.275508, URL http://doi.acm.org/10.1145/275487.275508

Serjantov A, Danezis G (2003) Towards an information theoretic metric for anonymity. In: Proceedings of the 2nd International Conference on Privacy Enhancing Technologies, Springer-Verlag, Berlin, Heidelberg, PET'02, pp 41–53, URL http://dl.acm.org/citation.cfm?id=1765299.1765303

Tanimoto SL, Itai A, Rodeh M (1978) Some matching problems for bipartite graphs. J ACM 25(4):517–525, DOI 10.1145/322092.322093, URL http://doi.acm.org/10.1145/322092.322093

Terrovitis M (2011) Privacy preservation in the dissemination of location data. SIGKDD Explor Newsl 13(1):6–18, DOI 10.1145/2031331.2031334, URL http://doi.acm.org/10.1145/2031331.2031334

Terrovitis M, Mamoulis N (2008) Privacy preservation in the publication of trajectories. In: Proceedings of the The Ninth International Conference on Mobile Data Management, IEEE Computer Society, Washington, DC, USA, MDM '08, pp 65–72, DOI 10.1109/MDM.2008.29, URL https://doi.org/10.1109/MDM.2008.29

Xiao Y, Xiong L (2015) Protecting locations with differential privacy under temporal correlations. In: Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, ACM, New York, NY, USA, CCS '15, pp 1298–1309, DOI 10.1145/2810103.2813640, URL http://doi.acm.org/10.1145/2810103.2813640

Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y (2010) T-drive: Driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA, GIS '10, pp 99–108, DOI 10.1145/1869790.1869807, URL http://doi.acm.org/10.1145/1869790.1869807

Yuan J, Zheng Y, Xie X, Sun G (2011) Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '11, pp 316–324, DOI 10.1145/2020408.2020462, URL http://doi.acm.org/10.1145/2020408.2020462

Zang H, Bolot J (2011) 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, ACM, New York, NY, USA, MobiCom '11, pp 145–156, DOI 10.1145/2030613.2030630, URL http://doi.acm.org/10.1145/2030613.2030630

Zhou B, Pei J, Luk W (2008) A brief survey on anonymization techniques for privacy preserving publishing of social network data. SIGKDD Explor

Newsl 10(2):12–22, DOI 10.1145/1540276.1540279, URL http://doi.acm.org/10.1145/1540276.1540279