Research Highlights (Required)

- Definition of sufficient sanitizer to protect privacy with utility guarantees.
- SwapMob algorithm is a sufficient sanitizer for real-time mobility data collection.
- Application to obtain Origin-Destination matrices that preserve privacy.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Swapping trajectories with a sufficient sanitizer

Julián Salas[a,b,**], David Megías[a,b], Vicenç Torra[c,d], Marina Toger[e], Joel Dahne[f], Raazesh Sainudiin[f,g]

[a] *Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC), Barcelona, Spain.*
[b] *CYBERCAT-Center for Cybersecurity Research of Catalonia, Barcelona, Spain.*
[c] *Hamilton Institute, Maynooth University, Maynooth, Ireland.*
[d] *School of Informatics, University of Skövde, Skövde, Sweden.*
[e] *Department of Economic and Cultural Geography, Uppsala University, Uppsala, Sweden.*
[f] *Department of Mathematics, Uppsala University, Uppsala, Sweden.*
[g] *Combient Competence Centre for Data Engineering Sciences, Uppsala University, Sweden.*

ABSTRACT

Real-time mobility data is useful for several applications such as planning transports in metropolitan areas or localizing services in towns. However, if such data is collected without any privacy protection it may reveal sensible locations and pose safety risks to an individual associated to it. Thus, mobility data must be anonymized preferably at the time of collection. In this paper, we consider the SwapMob algorithm that mitigates privacy risks by swapping partial trajectories. We formalize the concept of sufficient sanitizer and show that the SwapMob algorithm 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. This may be used for personalized assistants taking advantage of users' locations, so they can ensure user privacy while providing accurate response to the user requirements. We measure the privacy provided by SwapMob as the Adversary Information Gain, which measures the capability of an adversary to leverage his knowledge of exact data points 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.

1. Introduction

An increasing amount of location data is obtained from GPS, GSM and RFID technologies that may be integrated to our personal devices, such as our smartphones. This yields the opportunity of developing Location Based Services (LBS) that deliver content depending on users' locations.

However, revealing users' locations may have some privacy risks. If the data is linked to their real identities it may reveal personal preferences (e.g., sexual, political or religious orientation), or it may be used for inferring habits and knowing 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, cf. Terrovitis (2011).

We may consider privacy of location data when data is protected after it has been collected, or when users make location based queries. In the second case, privacy should be provided for location-based services (LBS), in such a way that the users' privacy is protected from service providers.

Beresford and Stajano (2003) proposed *mix zones* for privacy protection in LBS. Mix zones are regions in which applications cannot trace user movements. Inside a mix zone, applications receive users' pseudonyms assigned by a trusted intermediary that exchanges them when the users enter and exit the mix zone.

A similar method named SwapMob, based on the idea of interchanging pseudonyms but depending on proximity thresholds for location and time was proposed in Salas et al. (2018). This approach can be used for personalized assistants taking advantage of users' locations in real time, so they can ensure user privacy while providing accurate responses to the user requirements (instead of e.g. cloaking position to service providers), with the corresponding trade-off between privacy and utility.

**Corresponding author: Tel.: +34-933-263-651
e-mail: jsalaspi@uoc.edu (Julián Salas)

Considering the privacy and utility tradeoff, sanitization can be carried out by first setting a privacy parameter while preserving maximal possible utility (as in k-anonymity or ϵ-differential privacy) or can be carried out by setting first a utility parameter, while preserving maximal possible privacy. We consider that the sufficient statistics of the data are a good generic measure of utility. By preserving the sufficient statistics for a model after sanitization, we guarantee that any decision based on the estimated probability model of the data will be the same as without sanitization.

In this paper, we formalize the concept of sufficent sanitizers and apply it to the SwapMob algorithm to sanitize data at collection time. Then, we consider the specific case where the sufficient statistics to be preserved are the Origin-Destination matrices after we apply the SwapMob sanitizer.

The rest of the paper is organized as follows. In Section 2, we formalize the concept of sufficient sanitizer and provide some examples of sufficient statistcs related to traffic engineering and transportation planning. In Section 3, we recall the definition of SwapMob algorithm and show that it is a sufficient sanitizer. In Section 4, we evaluate our method with the additional utility guarantee of preserving Origin-Destination matrices. We finish with some conclusions and future work on Section 6.

2. Sufficient Sanitizers

In this section, we define sufficient sanitizers and provide examples of sufficient statistics for different statistical models. Sufficient santizers guarantee that the sufficient statistics after sanitization are not modified and, hence, the decisions obtained after sanitization will be equal to those based on an estimated probability model of the data before sanitization.

A *sanitizer* $\mathcal{S}$ is a randomized algorithm from the data space $\mathbb{D}$ to a sanitized data space $\widetilde{\mathbb{D}}$, i.e., $\mathcal{S}(d) = \widetilde{d} : \mathbb{D} \to \widetilde{\mathbb{D}}$, in such a way that the sanitized data in $\widetilde{\mathbb{D}}$ has additional privacy guarantees than in $\mathbb{D}$. For example, we may consider that $\widetilde{\mathbb{D}}$ is k-anonymous, or ϵ-differentially private. 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 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 for increasingly complex probability models 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. Most of the inference theoretic results we use are classical and can be found for example in Billingsley (1961).

2.1. State Counts

One of the simplest statistical experiments for mobility data is 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 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 the i-th state.

2.2. 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 than 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.

2.3. 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 matrices, infrastructure network capacity and traffic controls. OD trip generation models serve as a 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 under a given time interval. Such an ODM, as shown in Table 1 is a matrix of size $(n+1) \times (n+1)$ containing flow values N_{ij}, such as the 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). Note that a sequence of ODMs over a finite partition of time, say every hour in a typical weekday, are the sufficient statistics for a time-inhomogeneous Markov chain model over the n states and the 24 time steps. Moreover, an ODM-preserving sanitizer will produce sanitized trajectories that preserve the observed ODM and can thus be used for subsequent decisions.

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*.

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

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 that can preserve the ODMs from trajectory data will provide the utility from ODMs while ensuring additional privacy guarantees to the individuals associated with the trajectories.

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 4.3, we empirically assess the loss of privacy under a given metric as this spatial resolution varies for a single ODM (e.g., by Traffic Analysis Zones, ZIP code areas, square grid, etc.). Spatial aggregation of Origins and Destinations by zones can provide zone measurements and disaggregation by links can provide link-based counts. In other words, keeping overall ODM counts but not keeping the trajectory data in between can be used as input to traditional traffic allocation models. Keeping link flows but not keeping the OD for each trajectory enables to calibrate flows within these models.

2.4. 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 (e.g. 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.

3. SwapMob algorithm

In this section we provide some examples of how data can be used for making mobility maps and predictions that may be useful for intelligent transportation systems and for planning in a city. Then, we show that the SwapMob algorithm from Salas et al. (2018) is a sufficient sanitizer as it can preserve the sufficient statistics for the first three probability models of the previous section and some of their natural generalizations. Thus, it can be used to sanitize data for intelligent transportation systems and city planning.

Next, we give a high-level description of SwapMob in Algorithm 1. For a more detailed explanation of the algorithm please refer to Salas et al. (2018). SwapMob uses a grid partition (χ_k) of the space into squares of side-length χ and a partition (τ_j) of time in intervals of length τ. Users communicate their location data in real time to SwapMob sanitizer that swaps their IDs when they are co-located (with respect to χ and τ) and communicates their change of locations to the Service Provider. Therefore, the transitions between adjacent cells are kept intact, thus preserving the sufficient statistics of the Markov model over the partitions specified by the parameters: χ and τ.

Algorithm 1: SwapMob Algorithm from Salas et al. (2018)

Input: Trajectory database $d = (T_i)$ and partitions (τ_j) and (χ_k)
Output: Swapped trajectories $\widetilde{d} = \mathcal{S}(d)$
Initialization: $\widetilde{d} \leftarrow d$
for *each time interval* τ_j **do**
  for *each cell* χ_k **do**
    if *user* x *is located at cell* χ_k *at timestamp* $t_x \in \tau_j$ **then**
      add user x to swap list S_{τ_j,χ_k}
    end
  end
  Update $\widetilde{d}$ by swapping partial trajectories (T_x) between users $x \in S_{\tau_j,\chi_k}$
end
return *Swapped trajectories* $\widetilde{d}$

3.1. Use cases examples

In this section, we present three possible cases in which the SwapMob algorithm can be used to collect and sanitize data from users' sensors to protect it before processing.

First, Hoh and Gruteser (2005) propose that pre-specified vehicles 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.

Second, Calabrese et al. (2011) present a real-time urban monitoring platform and its application to the City of Rome. They use 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.

Third, Yuan et al. (2011) represent the knowledge from taxi-drivers as a landmark graph. 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 predictions and for providing a personalized routing service. In all three previous cases, SwapMob can be applied when collecting the data.

3.2. 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 SwapMob preserves the sufficient statistics of counts, transition counts and ODM: the three sufficient statistics with their corresponding statistical experiments described in Section 2.

It is easy to see that for a given time interval τ and spatial resolution specified by χ, the SwapMob sanitizer preserves the sufficient statistics of counts in each cell or state given by the τ,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 are 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 discussed 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.

Finally, note that any protection method that modifies the cell counts, transition counts or ODMs will not preserve the corresponding sufficient statistics of increasingly sophisticated probability models. If we add noise naively to locations or timestamps for ϵ-differential privacy or use averages for k-anonymity, then cell counts, transition counts and ODMs will be affected. Thus, naive ϵ-differential privacy and k-anonymity won't be sufficient sanitizers for these statistics.

4. Empirical evaluation

As shown in Section 3.2, SwapMob is a sufficient sanitizer for state counts and state transition counts, but it does not preserves the ODMs. In this section, we add constraints depending on the grid size of the Origin and Destination so that the ODMs are preserved. We define the adversary information gain to compare the privacy provided when preserving ODMs with different grid sizes.

We test 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.

4.1. 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 they are most likely not valid. We end up with 10,280 trajectories and 16,906,423 measurements.

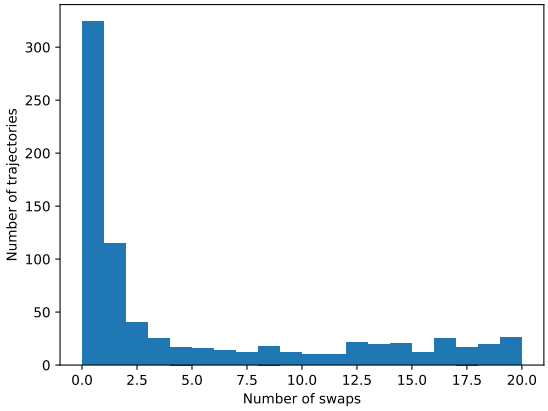


Fig. 1: Frequency histograms of number of swaps for trajectories participating in less than 20 swaps.

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 as one minute ($\tau = 60$), the same parameters for χ and τ used in Salas et al. (2018). These are about 6 and 3 times less than the average intervals for distance and time in the dataset. Other values of χ and τ may be assigned, however, this will change the precision of the data published and the number of swaps.

With these parameters, the number of possible swaps between all the trajectories is 641,262. The average number of swaps for each trajectory is 137. Of all the 10,280 trajectories there are only 772 trajectories with less than 20 swaps, and their distribution is depicted in Figure 1. There are 324 trajectories that 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).

4.2. Privacy measure: Adversary Information Gain

We define the Adversary Information Gain measure for privacy by adapting the Sensitive Attribute Risk measure from Salas (2019). Sensitive Attribute Risk considers the fraction of the published attributes of an individual that is part of its original attributes. The Adversary Information Gain (AIG) of a user's trajectory is the length of the longest segment without swapping in proportion to the length of the entire trajectory. It is the fraction of the original trajectory that can be disclosed by an adversary who knows that a data point belongs to a user, considering that the adversary propagates the knowledge of such data point to the whole segment in-between swaps.

We plot the distribution function of such measure in Figure 2b with grid size key as 'None' (since we are not conditioning on preserving any ODM here). It 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.

4.3. 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 2. 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 (grid size in Fig. 2. This is in addition to the earlier requirements for allowing a swap (τ, χ).

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 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 empirically evaluate 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 2a 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 and then we have the numbers for grids of squares of the given height. The largest height used is 1 degree. 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 which is 10 times the proximity threshold for swappability. Thus, for grid size 1, the sufficient statistics for the trips between the four quadrants NE, NW, SE and SW of Beijing would be preserved, but not with a high precision, as it would result from a grid size 0.01. In that case, the sufficient statistics for the trips with same origin and destination up to 1.11 km would be preserved. Since the preserved privacy heavily depends on the number of possible swaps, we expect it to quickly decrease as the grid becomes finer, see Figure 2b.

If the grid is too fine almost no swaps occur and the SwapMob algorithm returns almost the original data and preserves no privacy as measured through AIG. For very coarse grids the privacy is still reduced but depending on the application it could still be considered acceptable. This illustrates the trade-off between utility and privacy. If we decrease the grid size, the ODMs are more precise, but at the same time it is less likely that two individuals have the same origin and destination, thus it is less likely that there are possible swaps and therefore the resulting privacy decreases.

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 $[t_{i,o}, t'_{i,o}]$ and $[t_{i,d}, t'_{i,d}]$ 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 involving traffic flows. 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.

5. Related work

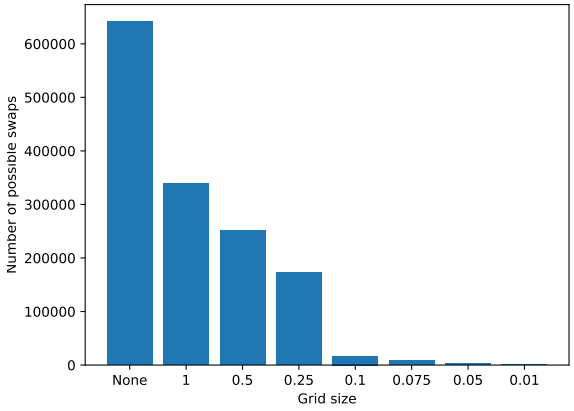
In this section we discuss some of the most relevant solutions for trajectory and location privacy. Hoh and Gruteser (2005) and Hoh et al. (2006) discuss the use of mobility data for transportation planning and traffic monitoring applications to provide drivers with feedback on road and traffic conditions. 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 who is periodically reporting his anonymized location information.

Hoh et al. (2006) consider the attack of deducing home locations of users by using clustering heuristics together with the decrease of speed reported by GPS sensors. Then, propose data suppression techniques by changing the sampling rate (e.g., from 1 to 2, 4 and 10 minutes) for protecting from such inferences.

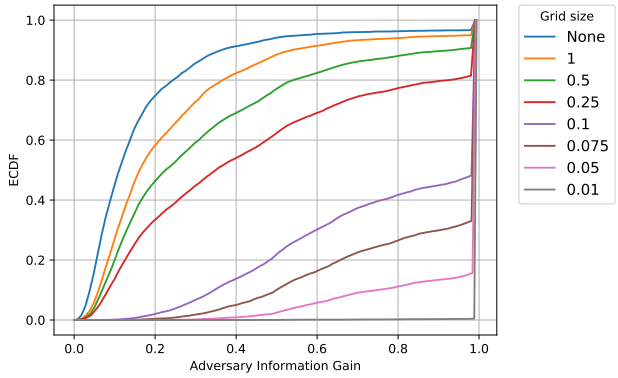
Hoh and Gruteser (2005) propose an algorithm to prevent adversaries from tracking complete individual paths, it perturbs slightly the trajectories of different individuals in such a way that multi-target tracking algorithms are not able to distinguish which segment of the path corresponds to which user. 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.

Mix Zones, introduced by Beresford and Stajano (2003), also prevent applications from tracking complete individual paths. They are used to preserve the advantages of location aware services while hiding users' identities from applications that receive their locations. 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. They 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.

To measure the location privacy provided by a mix zone, 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



(a) The number of possible swaps depending on the grid size.



(b) Empirical Cummulative Distribution Functions of the Adversary Information Gain for different grid sizes.

Fig. 2: Preserved privacy when also preserving the ODM

Serjantov and Danezis (2003) proposed for anonymous communications.

This is modeled by Beresford and Stajano (2004) as a movement matrix that represents the frequency of ingress and egress points to the mix zone at several times. 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.

An approach that does not consider middleware to obtain location privacy is proposed by Gidófalvi (2007, Chapter 9). It consists in 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 in 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 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 $\widetilde{\mathbb{D}}$ 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). 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). Their algorithm considers temporal correlations modeled as a Markov chain.

There are other techniques for anonymizing trajectories in data publishing and specifically for location privacy. For surveys on this topic cf. Fiore et al. (2019), Primault et al. (2019). For a more general overview on data privacy and big data technologies cf. Torra (2017).

6. Conclusions

We have defined the concept of sufficient sanitizer in which the utility requirement (as a sufficient statistic) is a priori defined, then the sanitization algorithm that preserves such utility is applied to the data.

We have shown that the SwapMob algorithm is a sufficient sanitizer for counts, transition counts and may be modified to preserve also ODMs. When applied in real time, it may be useful for providing anonymous data to personalized assistants. We have tested the SwapMob algorithm on the T-drive dataset and defined the Adversary Information Gain (AIG) measure to compare the privacy provided when using different grid sizes. AIG measures the capability of an adversary who knows exact points of the trajectory to infer a larger part of the full trajectory. We have added constraints on SwapMob to preserve the ODMs and performed experiments to show how AIG increases when we decrease the grid size for obtaining more precise ODMs. This is the natural tradeoff between the *societal utility* gained through the preservation of the ODM, where ODM is a sufficient statistic, and the *individual privacy lost* by the sufficient sanitizer. We remark that preserving sufficient statistics for various statistical decision problems is useful in traffic engineering and city planning, including exact count queries, transition count queries and ODM queries, which neither k-anonymity nor differential privacy can formally guarantee.

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 in a distributed computational setting to handle massive mobility data needs further investigations.

Acknowledgements

This work is partly funded by the Spanish Government through grants RTI2018-095094-B-C22 "CONSENT" and TIN2014-57364-C2-2-R "SMARTGLACIS", Swedish VR (project VR 2016-03346). Raaz Sainudiin was partly funded by Combient Competence Centre for Data Engineering Sciences at Uppsala University and the Research Center for Cyber Security at Tel Aviv University established by the State of Israel, the Prime Minister's Office and Tel-Aviv University. Julián Salas acknowledges the support of a UOC postdoctoral fellowship.

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, pp. 376–385.

Agrawal, R., Srikant, R., 1995. Mining sequential patterns, in: Proceedings of the Eleventh International Conference on Data Engineering, pp. 3–14.

Beresford, A.R., Stajano, F., 2003. Location privacy in pervasive computing. IEEE Pervasive Computing 2, 46–55.

Beresford, A.R., 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.

Billingsley, P., 1961. Statistical methods in markov chains. Ann. Math. Statist. 32, 12–40. doi:10.1214/aoms/1177705136.

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, 141–151.

Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C., 2013. Broadening the scope of differential privacy using metrics, in: Privacy Enhancing Technologies, pp. 82–102.

Chen, R., Fung, B.C., Desai, B.C., Sossou, N.M., 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, pp. 213–221.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography, pp. 265–284.

Evans, A., 1970. Some properties of trip distribution methods. Transportation Research 4, 19 – 36.

Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D.L., Aivodji, U.M., Olivier, B., Quertier, T., Stanica, R., 2019. Privacy of trajectory micro-data : a survey. CoRR abs/1903.12211.

Gidófalvi, G., 2007. Spatio-Temporal Data Mining for Location-Based Services. Ph.D. thesis. Faculties of Engineering, Science and Medicine Aalborg University. Denmark.

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, pp. 194–205.

Hoh, B., Gruteser, M., Xiong, H., Alrabady, A., 2006. Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Computing 5, 38–46.

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40, 63–74.

Kifer, D., Machanavajjhala, A., 2011. No free lunch in data privacy, in: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pp. 193–204.

Primault, V., Boutet, A., Mokhtar, S.B., Brunie, L., 2019. The long road to computational location privacy: A survey. IEEE Communications Surveys Tutorials 21, 2772–2793.

Reid, D.B., 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, 123–128.

Rodrigue, J.P., 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: Data Privacy Management, DPM 2015. Springer International Publishing, Cham, pp. 245–253.

Romero-Tris, C., Megías, D., 2018. Protecting privacy in trajectories with a user-centric approach. ACM Trans. Knowl. Discov. Data 12, 67:1–67:27.

Salas, J., 2019. Sanitizing and measuring privacy of large sparse datasets for recommender systems. Journal of Ambient Intelligence and Humanized Computing .

Salas, J., Domingo-Ferrer, J., 2018. Some basics on privacy techniques, anonymization and their big data challenges. Mathematics in Computer Science 12, 263–274.

Salas, J., Megías, D., Torra, V., 2018. Swapmob: Swapping trajectories for mobility anonymization, in: Privacy in Statistical Databases, Springer International Publishing, Cham. pp. 331–346.

Salas, J., Torra, V., 2018. A general algorithm for k-anonymity on dynamic databases, in: Data Privacy Management, Cryptocurrencies and Blockchain Technology, Springer International Publishing, Cham. pp. 407–414.

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, p. 188.

Serjantov, A., Danezis, G., 2003. Towards an information theoretic metric for anonymity, in: Proceedings of the 2nd International Conference on Privacy Enhancing Technologies, pp. 41–53.

Tanimoto, S.L., Itai, A., Rodeh, M., 1978. Some matching problems for bipartite graphs. J. ACM 25, 517–525.

Terrovitis, M., 2011. Privacy preservation in the dissemination of location data. SIGKDD Explor. Newsl. 13, 6–18.

Torra, V., 2017. Data privacy: Foundations, new developments and the big data challenge. Springer.

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, pp. 1298–1309.

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, pp. 316–324.

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, pp. 99–108.