Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Julián Salas, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature_____ Date_____

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

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.
Justification for re-publication: We define and explore the concept of sufficient sanitizer and apply it to SwapMob algorithm that is defined in the published conference paper.

Research Highlights (Required)

- Definition of the concept of sufficient sanitizer for preserving privacy and utility.
- SwapMob algorithm is a sufficient sanitizer for different sufficient statistics.
- Tested the application to obtain Origin-Destination matrices with privacy guarantees.

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 anonymization approach SwapMob 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.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction and related work

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 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).

In this paper we use the SwapMob algorithm from Salas et al. (2018). This approach for anonymization of trajectories can be used for personalized assistants taking advantage of users' locations in real time, so they can ensure user privacy while providing accurate response to the user requirements (instead of e.g. cloaking position to service providers).

Different solutions have been proposed for anonymizing trajectories in data publishing, cf. Fiore et al. (2019). Terrovitis and Mamoulis (2008) consider a discrete spatial domain were the user trajectories are expressed as sequences of points of interest (POIs).

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 mon-

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

itoring 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.

To prevent adversaries from tracking complete individual paths, Hoh and Gruteser (2005) propose an algorithm that perturbs slightly the trajectories of different individuals in such a way so 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.

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.

They design a model for location privacy protection that aims to preserve the advantages of location aware services while hiding users' identities from applications that receive their locations.

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.

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. 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 location-based services (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 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 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 $\tilde{\mathcal{D}}$ satisfying ϵ-differential privacy. For measuring utility, in the STM case, sanitized data are mainly used to perform two data mining tasks, count query and frequent sequential pattern mining (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, 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 formalize the concept of sufficent sanitizers and apply it to the SwapMob algorithm to anonymize data at collection time, which may be used for personalized assistants taking advantage of users' locations. We test the SwapMob sanitizer when the sufficient statistic to be preserved are the Origin-Destination matrices.

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 also when including the additional utility guarantee of preserving Origin-Destination matrices. We finish with some conclusions and future work on Section 5.

2. Sufficient Sanitizers

Considering the trade-off between privacy and utility, there are two approaches for data protection: (i) the privacy first approach in which the data is released based on a privacy parameter as in k-anonymity or ϵ-differential privacy, and (ii) the utility first approach whereby a required utility is set and the algorithm is guaranteed to provide the utility while preserving maximal possible privacy. In this section, we take the second approach and define sufficient sanitizers that provide the utility from any decision based on an estimated probability model of the data by preserving the corresponding sufficient statistics.

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 $\mathbb{D}$.

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 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 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 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. 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).

ODM are constructed based on estimations from travel studies as part of traffic census: field, online and telephone traffic

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

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 will keep the utility from ODM while providing 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 show that the algorithm SwapMob from Salas et al. (2018) is a sufficient sanitizer.

We recall that SwapMob may be used for sanitization during data collection. The individuals communicate their data in real time to the SwapMob anonymizer that removes their IDs and communicates their locations to the Service Provider, which communicates back to SwapMob anonymizer when there are records that are closer than the prespecified thresholds for time τ and location χ, to let SwapMob swap their IDs. In this way, SwapMob exchanges segments of trajectories that are near based on proximity thresholds for time τ and location χ. See Salas et al. (2018) for the detailed protocol.

3.1. Utility of swapped data as privacy-preserving decisions

In this section and the entire paper, we are assuming that the data sanitized by SwapMob will be used for making mobility maps and predictions that may be useful for intelligent transportation systems and for planning in a city.

3.1.1. 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.

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.

The offline mining from Yuan et al. (2011) representing the knowledge from taxi-drivers as a landmark graph may still be carried out after data anonymization with SwapMob. 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.

3.1.2. 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 ODM: the three sufficient statistics with their corresponding statistical experiments described in Section 2.

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 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.

4. 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.

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 the measurement is most likely not valid. 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 these parameteres, 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, 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

For measuring privacy we will define the Adversary Information Gain measure 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.

By considering the sensitive attributes to be locations instead of movies in recommender systems, the Adversary Information Gain (AIG) here is the fraction of the original trajectory that is disclosed by an adversary who knows a point of it. 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 measurement forward and backward in time until the next and previous swaps occured. 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 2b with grid size = None. 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.

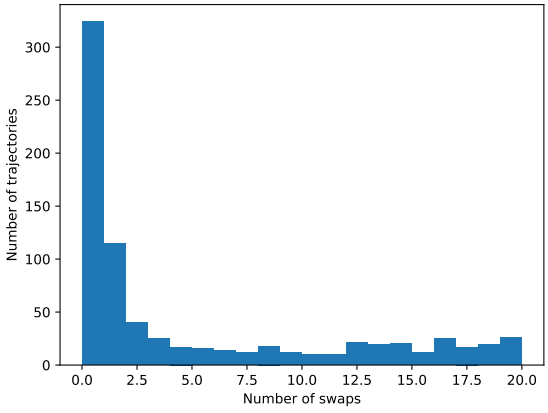


Fig. 1: Frequency histograms of number of swaps for trajectories participating in less than 20 swaps.

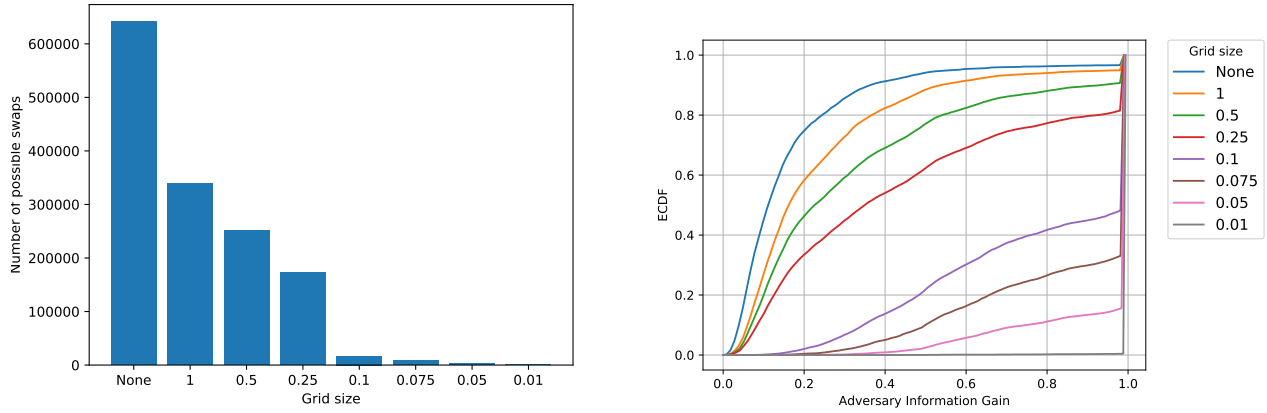
4.3. Privacy when preserving the Origin-Destination Matrix

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). 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 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 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



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

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

Fig. 2: Preserved privacy when also preserving the ODM

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. We see that the number of possible swaps quickly decreases as the grid gets finer.

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.

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 $[t_{i,o}, t'_{i,o}]$ 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. 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. Conclusions

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

We showed that the SwapMob algorithm is a sufficient sanitizer. When applied in real time it may be useful for providing anonymous data to personalized assistants.

We tested the SwapMob algorithm on the T-drive dataset to show that a high number of swaps occur and most trajectories participate in several swaps. The original trajectories are split up into many smaller segments, hence SwapMob prevents an adversary who may know exact points of the trajectory from inferring much of the full trajectory, this is measured with the Adversary Information Gain.

However, when adding the utility restriction of preserving the ODM, the number of swaps quickly decreases and it becomes much easier for an adversary to infer a bigger portions of the trajectories in the database.

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 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.

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 grants RTI2018-095094-B-C22 "CONSENT" and TIN2014-57364-C2-2-R "SMART-GLACIS", 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. pp. 376–385.

Agrawal, R., Srikant, R., 1995. Mining sequential patterns, in: Proceedings of the Eleventh International Conference on Data Engineering, IEEE Computer Society, Washington, DC, USA. 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: De Cristofaro, E., Wright, M. (Eds.), Privacy Enhancing Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg. 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, ACM, New York, NY, USA. pp. 213–221.

Damiani, M.L., 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. pp. 32–41.

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, 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.

Ghinita, G., Kalnis, P., Kantarcioglu, M., Bertino, E., 2011. Approximate and exact hybrid algorithms for private nearest-neighbor queries with database protection. GeoInformatica 15, 699–726.

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, IEEE Computer Society, Washington, DC, USA. 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, ACM, New York, NY, USA. pp. 193–204.

Paulet, R., Kaosar, M.G., Yi, X., Bertino, E., 2014. Privacy-preserving and content-protecting location based queries. IEEE Trans. on Knowl. and Data Eng. 26, 1200–1210.

Pensa, R.G., Monreale, A., Pinelli, F., Pedreschi, D., 2008. Pattern-preserving k-anonymization of sequences and its application to mobility data mining, in: PiLBA.

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: 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.

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: Domingo-Ferrer, J., Montes, F. (Eds.), 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: 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.

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. pp. 188–.

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. 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.

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. pp. 65–72.

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. 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, ACM, New York, NY, USA. 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, ACM, New York, NY, USA. pp. 99–108.