P

Privacy of Aggregated Mobility Data

Gergely Acs, Szilvia Lestyán, and Gergely Biczók
CrySyS Lab, Department of Networked Systems and Services, Budapest University of Technology and Economics (BME-HIT), Budapest, Hungary

Synonyms

Anonymization of aggregated location information; Anonymization of aggregated mobility data; Privacy of aggregated location information

Definitions

Aggregated mobility data is a function of the number of individuals visiting a given set of locations over a given time of interest. A *location* refers to a well-separated geographical region.

Background

Mobility datasets are invaluable in traffic management, service accessibility and in understanding complex social processes such as the spreading of diseases or the exchange of information among individuals. As personal mobility patterns reveal tremendous sensitive information about individuals, publishing and sharing mobility datasets could harm their privacy. Moreover, even being part of a (mobility) dataset could reveal sensitive personal information if the aggregate relate to a group of users sharing a sensitive characteristic, e.g., being infected with a virus. One might argue that publishing aggregate information, such as the number of individuals at a given location, is enough to reconstruct aggregate mobility patterns and has no privacy implications. However, this reasoning is flawed as evidenced by a handful of attacks reported in the literature.

Differencing attacks work against counting queries executed on the location trajectories. The querier is interested in the number of people whose trajectories satisfy a specified condition (e.g., the number of trajectories which contain a certain hospital). Queries can be filtered instantly by an auditor, e.g., all queries which have too small support, say less than k (i.e., only k trajectories satisfy the condition), are simply refused to answer. However, this approach is not enough to prevent privacy breaches; if the support of two queries are both greater than k, their difference can still be 1. Defenses against such attacks are not straightforward, e.g., verifying whether the answers of two or more queries disclose any location visit can be computationally infeasible.

The attack described in Xu et al. (2017) successfully reconstructed more than 70% of 100,000 trajectories merely from the total number of visits at 8000 locations, which were published

S. Jajodia et al. (eds.), *Encyclopedia of Cryptography, Security and Privacy*,
https://doi.org/10.1007/978-3-642-27739-9_1575-1

every half an hour over a whole week in a large city. The attack exploits three fundamental properties of location trajectories: predictability (the current location can be well-predicted from the previous location), regularity (most people visit very similar locations every day), and uniqueness (every person's mobility pattern is sufficiently dissimilar to others'). The first phase of the attack exploits predictability and reconstructs every trajectory within every single day. The second phase exploits regularity and uniqueness and reconstructs complete trajectories by identifying their daily fragments. Finally, the last phase reidentifies individuals using the uniqueness property again: a few locations of any individual known from external sources (e.g., social media) will single out the individual's trajectory (De Montjoye et al. 2013).

The attack described in Pyrgelis et al. (2018) was successful in membership inference, i.e., determining whether or not the data of a target user was part of location aggregates. Focusing on distinguishing between location aggregates that include the data of the target user from those that do not, the attack aims to infer the target's membership in unseen aggregate statistics by training a machine learning classifier on the prior knowledge of the adversary (past users' locations, aggregates of groups including and excluding the target user). Results show that releasing raw aggregates poses a significant privacy threat, especially if the adversary knows the location of a small group of user including the target or when it has prior information on user groups on which it carries out the membership inference.

Consequently, aggregation per se does not necessarily prevent privacy breaches, and additional countermeasures are needed to guarantee privacy for individuals even in a dataset of aggregate mobility data such as spatiotemporal densities.

Theory

Suppose a geographical region which is composed of a set $\mathbb{L}$ of locations visited by N individuals over a time of interest with T discretized epochs (an epoch can be any time interval such as a second, a minute, an hour, etc.). These locations may represent a partitioning of the region (e.g., all districts of the metropolitan area of a city). The mobility dataset D of N users is a binary data cube with size $N \cdot |\mathbb{L}| \cdot T$, where $D_{i,L,t} = 1$ if individual i visited location L in epoch t otherwise $D_{i,L,t} = 0$. That is, each individual's record (or trajectory) is represented by a binary vector with size $|\mathbb{L}| \times T$. The spatiotemporal density of locations $\mathbb{L}$ is defined by the number of individuals who visited these locations as a function of time. More precisely, there is a time series $\mathbf{X}^L = \langle X_0^L, X_1^L, \ldots, X_{T-1}^L \rangle$ for any location $L \in \mathbb{L}$, where $X_t^L = \sum_{i=1}^N D_{i,L,t}$ and $0 \leq t < T$. $\mathbf{X}^{\mathbb{L}}$ denotes the set of time series of all locations $\mathbb{L}$ and is referred to as the spatiotemporal density of locations $\mathbb{L}$ in the sequel.

In general, any data release, including that of any aggregated mobility data, is modeled by releasing the results of data queries. For example, if the querier is interested in the spatiotemporal density of locations $S_L \subseteq \mathbb{L}$ at time $S_T \subseteq \{0, 1, \ldots, T-1\}$, then the query $Q(S_L, S_T)$ is computed as $Q(S_L, S_T) = \sum_{L \in S_L, t \in S_T} \sum_{i=1}^N D_{i,L,t} = \sum_{L \in S_L, t \in S_T} X_t^L$. There are at least three approaches for the privacy-preserving release of aggregated mobility data:

Approach 1: Anonymization of specific query results – compute any query Q on the original data D (or $\mathbf{X}^{\mathbb{L}}$) and release only the anonymized query result $\hat{Q}(S_L, S_T)$;

Approach 2: Anonymization of the mobility dataset – anonymize the mobility dataset D into $\hat{D}$, then release $\hat{D}$ which can be used to answer any query Q as $\hat{Q}(S_L, S_T) = \sum_{L \in S_L, t \in S_T} \sum_{i=1}^N \hat{D}_{i,L,t}$;

Approach 3: Anonymization of spatiotemporal density – compute the density $\mathbf{X}^{\mathbb{L}}$ from the original mobility data D as $X_t^L = \sum_{i=1}^N D_{i,L,t}$, and release the anonymized $\hat{\mathbf{X}}^{\mathbb{L}}$, where $\hat{\mathbf{X}}^{\mathbb{L}}$ can be used to answer any query Q.

In Approach 1, a querier can adaptively (i.e., interactively) choose its queries depending on the result of previously answered queries. By contrast, in Approaches 2 and 3, the released data

are used to answer arbitrary number and type of queries noninteractively (i.e., the queries are independent of each other).

Approach 1: Anonymization of Specific Query Results

Syntactic Anonymization

Privacy breaches may be alleviated by query auditing which requires to maintain all released queries. The database receives a set of counting queries $Q_1(S_{L_1}, S_{T_1}), \ldots, Q_n(S_{L_n}, S_{T_n})$, and the auditor needs to decide whether the queries can be answered without revealing any single visit or not. Specifically, the goal is to prevent the *full disclosure* of any single visit of any spatiotemporal point in the dataset.

Definition 1 (Full disclosure) $D_{i,L,t}$ is fully disclosed by a query set $\{Q_1(S_{L_1}, S_{T_1}), \ldots, Q_n(S_{L_n}, S_{T_n})\}$ if $D_{i,L,t}$ can be uniquely determined, i.e., in all possible datasets D consistent with the answers $\mathbf{c} = (c_1, \ldots, c_n)$ to queries $Q_1, \ldots, Q_n$, $D_{i,L,t}$ is the same.

As each query corresponds to a linear equation on location visits, the auditor can check whether any location visit can be uniquely determined by solving a system of linear equations specified by the queries.

Anonymization with Differential Privacy

An alternative approach to query auditing perturbs each query result with some random noise and releases these noisy answers. The noise magnitude must be $\Omega(\sqrt{N})$ in order to have any reasonable privacy guarantee, where N is the number of trajectories in the dataset. For (ε, δ)-differential privacy, the added noise usually follows a Laplace or Gaussian distribution.

Approach 2: Anonymization of the Mobility Dataset

Syntactic Anonymization

In general, anonymizing location trajectories (i.e., the whole cube D) while preserving practically acceptable utility is challenging. This is due to the fact that location data is typically high-dimensional and sparse.

Most k-anonymization schemes generalize multiple trajectories into a single group (or cluster) with size at least k and represent each trajectory with the centroid of their cluster. Hence, every trajectory becomes (syntactically) indistinguishable from all other trajectories within its cluster.

Unfortunately, such approaches fail to provide sufficiently useful anonymized datasets because of the *curse of dimensionality*: any trajectory exhibits almost identical similarity to any other trajectory in the dataset. This implies that the centroid of each cluster tend to be very dissimilar from the cluster members implying weak utility.

To improve utility while relaxing privacy requirements, k^m-anonymity has also been considered to anonymize location trajectories, which requires that, for any m locations, there should exist 0 or at least k trajectories in the dataset containing them. However, most anonymization solutions guaranteeing k^m-anonymity have a computational cost which is exponential in m in the worst case, and hence they are only scalable to small values of m.

Anonymization with Differential Privacy

A more promising approach is to publish a synthetic (anonymized) mobility dataset resembling the original dataset as much as possible, while achieving provable guarantees w.r.t. the privacy of each individual. The records in both datasets follow similar underlying distributions, i.e., after modeling the generator distribution of the original dataset, random samples (records) are drawn from a noisy version of this distribution. A few solutions exist in literature where the generator distribution is modeled explicitly and noised to guarantee differentially privacy. For example, DP-WHERE (Mir et al. 2013) adds noise to the set of empirical probability distributions which is derived from CDR (call detail record) datasets and samples from these distributions to generate synthetic CDRs which are differential private.

Some other works generate synthetic sequential data using more general data-generating models such as different Markov models. Although

these approaches have wide applicability, they are usually not as accurate as a specific model (like above) tailored to the *publicly known* characteristics of the dataset to be anonymized.

Approach 3: Anonymization of Spatiotemporal Density

Publishing the time series $\mathbf{X}^{\mathbb{L}}$ is equivalent to releasing the results of $|\mathbb{L}| \times T$ queries over a time of interest with T epochs, where a query, which is specified with a given location L and epoch t, returns $X_t^L = \sum_{i=1}^{N} D_{i,t,L}$. Differential privacy is used in many practical scenarios where query results are interpreted as location counts.

Some works address the release of time series data with the guarantees of differential privacy. Most of these methods reduce the global sensitivity of the time series by using standard lossy compression techniques borrowed from signal processing such as sampling, low-pass filtering, Kalman filtering, and smoothing via averaging. The main idea is that the utility degradation is decomposed into a reconstruction error, which is due to lossy compression, and a perturbation error, which is due to the injected Laplace or Gaussian noise to guarantee differential privacy. Although strongly compressed data is less accurate, it also requires less noise to be added to guarantee privacy. The goal is to find a good balance between compression and perturbation to minimize the total error.

In the context of releasing multilocation traffic aggregates, road network and density are utilized to model the autocorrelation of individual regions over time as well as correlation between neighboring regions (Fan et al. 2013). Drawing on the notion of w-event privacy, RescueDP studies the problem of the real-time release of population statistics per regions. Such w-event privacy protects each user's mobility trace over any successive w time stamp inside the infinite data grouping algorithm that dynamically aggregates sparse regions together.

A practical scheme for releasing the spatiotemporal density of a large municipality based on a large CDR dataset is introduced in Acs and Castelluccia (2014). Differential privacy is guaranteed by adding Gaussian noise to the location counts. Several optimizations are applied to boost accuracy: time series are compressed by sampling, clustering, and low-pass filtering. The approach is demonstrated by anonymizing the spatiotemporal density of the city of Paris in France.

Open Problems and Future Directions

There are at least two interesting future directions to explore.

First, the data-generating distribution of human mobility can be implicitly modeled using generative artificial neural networks (ANNs) such as recurrent neural networks (RNNs). Generative ANNs have exhibited great progress recently, and their representational power has been demonstrated by generating very realistic (but still artificial) sequential data such as texts or music. The intuition is that, as deep ANNs can "automatically" model very complex data-generating distributions thanks to their hierarchical structure, they can potentially be used to produce realistic synthetic sequential data and eventually aggregated mobility data. To guarantee differential privacy for such synthetic aggregates, model construction (learning) should be perturbed with carefully calibrated (Gaussian) noise.

Second, current anonymization approaches release aggregated mobility data only over a limited time interval. To release data over a longer period, one has to use a privacy model that supports composition (e.g., differential privacy). Finding the optimal (tightest) bound of the privacy guarantee of the composition of multiple releases has been an active research field.

Cross-References

- ▸ Adversarial/External Knowledge (Privacy in the Presence of)
- ▸ Anonymity
- ▸ Differential Privacy
- ▸ Inference Control

- k-Anonymity
- Location Information (Privacy of)
- Location Privacy
- Microdata Anonymization Techniques
- Spatiotemporal Privacy

References

Acs G, Castelluccia C (2014) A case study: privacy preserving release of spatio-temporal density in Paris. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)

De Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3:1376

Fan L, Xiong L, Sunderam V (2013) Differentially private multi-dimensional time series release for traffic monitoring. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer, pp 33–48s

Mir DJ, Isaacman S, Cáceres R, Martonosi M, Wright RN (2013) DP-WHERE: differentially private modeling of human mobility. In: BigData Conference, pp 580–588

Pyrgelis A, Troncoso C, Cristofaro ED (2018) Knock knock, who's there? Membership inference on aggregate location data. In: 25th Annual Network and Distributed System Security Symposium (NDSS)

Xu F, Tu Z, Li Y, Zhang P, Fu X, Jin D (2017) Trajectory recovery from ash: user privacy is not preserved in aggregated mobility data. In: Proceedings of the 26th International Conference on World Wide Web (WWW), pp 1241–1250