Computers & Security 97 (2020) 101953

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Aggregation-based location privacy: An information theoretic approach

Wenjing Zhang[a], Bo Jiang[b], Ming Li[b,*], Ravi Tandon[b], Qiao Liu[a], Hui Li[a]

[a] School of Cyber Engineering, Xidian University, Xi'an, Shaanxi, 710071, China
[b] Department of Electrical and Computer Engineering, University of Arizona, Tucson, 85721, AZ

ARTICLE INFO

Article history:
Received 31 December 2019
Revised 30 June 2020
Accepted 2 July 2020
Available online 22 July 2020

Keywords:
Privacy metric
Aggregation-based location privacy
Information-theoretic approach
Privacy-utility tradeoff
Upper bound
Rate distortion function

ABSTRACT

We explore the problem of quantifying and protecting aggregation-based location privacy and study the privacy-utility tradeoff, which are essential to protect user's location privacy when releasing aggregate statistics. Existing works on Aggregation-based Location Privacy Protection Mechanisms (ALPPMs) are mainly based on differential privacy, and metrics for evaluating information leakage introduced by releasing aggregates are normally built on adversary's estimation error. However, there lacks privacy metrics for measuring the fundamental leakage on individual user's data that is independent of specific data instances or attack algorithms. In this paper, we aim to solve this problem using an information-theoretic approach. We first propose a privacy metric based on the mutual information between the individual user's location profile and the released location aggregates, and formulate the optimal location aggregate release problem that minimizes the mutual information given a utility constraint for each user. Since solving this optimization problem causes exponential complexity, we turn to prove and evaluate an upper bound, i.e., the mutual information between the original and the perturbed location aggregates, and propose a Blahut-Arimoto based algorithm to solve the optimization problem minimizing the mutual information to derive an ALPPM. We validate the actual leakage of our ALPPM and compare it to a differentially private mechanism through experiments over both synthetic and real-world location datasets. Results show the advantage of the proposed ALPPM in terms of privacy-utility tradeoff, which is enhanced when users' location prior distributions are more skewed.

1. Introduction

Over the past few years, location information has enabled massive personalized services and intelligent applications, which makes our daily lives much more convenient than before. This motivates legal organizations and companies to collect more and more location data from many different sources to support location intelligence technologies. In many cases, they may only need to periodically compute and release the aggregate statistics calculated based on a large number of users' location profiles to identify certain phenomena or track important patterns. Examples of aggregate statistics could be summation, mean, standard deviation, density, and so on, depending on different purposes of services. For instance, the XData project publishes a cellular dataset which only reports the population density of each region, i.e., the number of users covered by a cellular tower at a certain timestamp, which is useful in identifying areas where to develop new businesses or build new infrastructures (Acs and Castelluccia, 2014). Other applications based on location aggregates include obtaining the average speed of road traffic during rush hours (Barth, 2009), calculating average delay on a road segment to avoid traffic congestion (Popa et al., 2011), and counting the number of users infected by a new flu at a certain location to monitor its propagation (Li et al., 2014). Undoubtedly, it's essential to use aggregate location statistics to improve people's lives.

Even though publishing location aggregates rather than users' original location data can protect their location privacy, previous works (Pyrgelis et al., 2017a; 2017b; Xu et al., 2017) have shown that aggregates still leak information about individuals' locations. Specifically, an adversary is capable of inferring the presence of a target user's location within a dataset (Pyrgelis et al., 2017a). Moreover, Pyrgelis et al. demonstrated that an adversary can accurately deduce a significant fraction of users' locations and mobility profiles from raw aggregates (Pyrgelis et al., 2017b). In addition, many users choose not to install applications which may log location aggregates from their mobile devices due to privacy concerns (Reid, 2011; Riley, 2008). Obviously, privacy concern has posed a major obstacle to releasing raw aggregates to the public as well as

* Corresponding author.
E-mail addresses: xd.zhangwenjing@gmail.com (W. Zhang), bjiang@email.arizona.edu (B. Jiang), lim@email.arizona.edu (M. Li), tandonr@email.arizona.edu (R. Tandon), qiaoliu@xidian.edu.cn (Q. Liu), lihui@mail.xidian.edu.cn (H. Li).

https://doi.org/10.1016/j.cose.2020.101953
0167-4048/

	t=1	t=2	t=3
l=1	0	0	2
l=2	2	0	0
l=3	0	2	0

(a) Aggregate location matrix $\boldsymbol{A}$.

	t=1	t=2	t=3
l=1	0	0	1
l=2	1	0	0
l=3	0	1	0

(b) User m's location profile $\boldsymbol{A}_m$.

Fig. 1. Inferring $\boldsymbol{A}_m$ from $\boldsymbol{A}$.

sharing them with third parties, meaning that they need to be perturbed before release to protect users' location privacy.

A motivating example is presented in Fig. 1. We consider a simple case where there are only two users participating in an aggregation process and a trusted aggregator is responsible for counting how many users visiting each location. When the trusted aggregator directly releases the original aggregate location matrix $\boldsymbol{A}$ to the public, a malicious adversary could infer a target user's location profile $\boldsymbol{A}_m$ from $\boldsymbol{A}$, which are illustrated in Fig. 1a and b. Here row l represents different locations and column t represents different timestamps, and each element is the count of occurrences of each location. A malicious adversary with the background knowledge that a target user m has participated in this location aggregation, could infer user m's location profile $\boldsymbol{A}_m$ accurately, since both users visited the same location at all timestamps. This is obviously a privacy violation to user m. As a result, in order to motivate more users to participate in location aggregation, the trusted aggregator should protect users' location profiles by releasing the aggregate location $\boldsymbol{A}$ in a privacy-preserving way.

1.1. Related work

1.1.1. Aggregation-based privacy protection

Most existing works on privacy-preserving aggregate statistics release are based on Differential Privacy (DP), which is an indistinguishability-based notion aiming to protect individual's data while releasing aggregate information about a database (Acs and Castelluccia, 2014; Chan et al., 2011; Dwork, 2008; Dwork et al., 2010; Ho and Ruan, 2011). For instance, Fan et al. proposed an adaptive scheme to release real-time location statistics under DP which provides improved utility (Fan and Xiong, 2012). Other privacy-preserving aggregate release mechanisms based on DP are proposed in Rastogi and Nath (2010) and Erlingsson et al. (2014), where the applications were to distributed time-series data and crowdsourcing statistics from end-user client software respectively. However, DP is context-free metric which considers the worst case adversary (who may possess arbitrary background knowledge and know every other users' data except the interested user's data), thus lead to worse utility-privacy tradeoff than context-aware notions, as shown by the work in Jiang et al. (2018) and Huang et al. (2017). In addition to DP, Li et al. leveraged additive homomorphic encryption and a novel key management technique to perform aggregation with high efficiency (Li et al., 2014), but this method can not be used to quantify information leakage.

1.1.2. Privacy metrics

Privacy metrics for quantifying individual location privacy (Oya et al., 2017; Shokri et al., 2011) or trace privacy (Zhang et al., 2019) were proposed in the literature, but they are not directly applicable to evaluate location aggregates. In addition, Pyrgelis et al. studied the feasibility of membership inference attacks in the context of aggregate location data (Pyrgelis et al., 2017a), and proposed to measure users' privacy loss from aggregate location time-series and evaluated the privacy protection levels offered by existing defense mechanisms based on DP (Pyrgelis et al., 2017b). However, no privacy protection mechanism has been proposed for location aggregates in these works. Even though Pyrgelis et al. evaluated the information leakage of individuals' punctual locations and mobility profiles introduced by the released aggregates (Pyrgelis et al., 2017b), their metrics in quantifying privacy via concrete inference algorithms or attacks have a limitation that the results may vary depending on different datasets or adversary's background information. However, we quantify and bound the amount of information about individuals' locations revealed from location aggregates independent of any dataset and background information.

There are also other set of privacy metrics reviewed in Wagner and Eckhoff (2018) and Primault et al. (2018), most of which are built on entropy, mutual information, k-anonymity (Sweeney, 2002), DP, and adversary's attack correctness (Shokri et al., 2011). Entropy isn't a proper metric for our problem setting since it doesn't capture the additional leakage introduced by the released data. K-anonymity and DP are originally proposed to protect the existence of a single record in a database, but k-anonymity was shown to suffer from various attacks with background information, and has been disregarded after the appearance of DP (Shokri et al., 2010). Shokri et al. proposed privacy metrics that quantify attacker's location estimation error under specific types of inference attacks, which inherently takes location correlations into account (Shokri et al., 2011). However, this metric assumes specific type of inference attacks, while information-theoretic metric is agnostic to specific attack algorithms.

1.1.3. Information theoretic privacy

Authors in a few works (Ma and Yau, 2015; Oya et al., 2017; du Pin Calmon and Fawaz, 2012; Sankar et al., 2013; Zhang et al., 2019) utilized information theoretic approaches, such as mutual information and conditional mutual information, to measure privacy leakage and designed privacy protection methods. The main advantage of such metrics is that they are context-aware, meaning that they consider the prior knowledge of data in the privacy definition and exploit it in mechanism design by adding noise selectively based on the data priors, so as to achieve a higher utility-privacy tradeoff. Specifically, the privacy metrics proposed in Ma and Yau (2015) and Oya et al. (2017) are only applicable to single location scenario. The privacy metrics proposed in Sankar et al. (2013) and du Pin Calmon and Fawaz (2012) have scalability issues if they are used on a large domain size, and are thus not practical to be applied to aggregation-based location privacy. In addition, Zhang et al. studied the problem of an individual user's location trace privacy while assuming the server is untrusted (Zhang et al., 2019). We want to highlight that our work in this paper is the first one to use context-aware metric to quantify and protect aggregation-based location privacy with a trusted aggregator who adds noise during data release to the public, making our problem setting become entirely different from the one in Zhang et al. (2019). Moreover, the problem formulation and solution approaches in this paper are more challenging, since the leakage involves multiple users instead of single user. Even though we all proposed algorithms by modifying the Blahut-Arimoto algorithm, the inputs and outputs of these two algorithms are completely different. Lastly, our works also have substantial differences in simulation setup and experimental results, each of which provides valuable insights in its own problem setting and has no resemblance to one another.

To sum up, the problems of quantifying privacy using information-theoretic measures and designing context-aware privacy mechanisms for aggregate location release in the centralized setting have not been well studied and we aim to solve these problems in this paper.

1.2. Contributions

The major contributions of this paper are summarized as follows:

- This is the first work to use context-aware metric to measure aggregation-based location privacy. Specifically, we propose a privacy metric based on mutual information to measure individual user's information leakage caused by releasing perturbed location aggregates to the public. In particular, we consider a special form of aggregate statistics, i.e., summation, which is commonly used in real-world scenarios. Then we formulate the optimal ALPPM as a minimization problem over the leakage given a utility constraint. The proposed metric is generic and independent of any specific inference attack, meaning that it can provide us a formal method of measuring and comparing the strength of privacy guarantees offered by different ALPPMs.
- We evaluate the information leakage on the original location aggregates introduced by the perturbed location aggregates, and formulate an optimization problem to derive an ALPPM. Due to the issue of exponential complexity, we prove an upper bound on the privacy-utility tradeoff for location aggregates and obtain the optimal ALPPM according this upper bound. Interestingly, the maximal individual user's leakage under this optimal ALPPM is proved to be a tighter upper bound on individual user's leakage.
- We compare the proposed ALPPM with a differentially private mechanism presented in Chan et al. (2011) based on the proposed privacy metric over both synthetic and real-world location datasets. Results show that our proposed ALPPM reveals less information under the same utility constraint and its advantage in the privacy-utility tradeoff becomes more conspicuous when users' location prior distributions are more skewed and there are more users with skewed priors. In addition, results also indicate that unpopular locations are better protected than popular locations under the proposed ALPPM.

Clearly, our work contribute to motivating people to take part in location aggregation, which is essential for generating valuable aggregate location datasets for statistical analysis while protecting user's location privacy.

We organize the rest of our paper as follows. The problem statement and preliminaries are presented in Section 2 and 3 respectively. Section 4 shows the main results of the privacy-utility tradeoff for location aggregates and also the algorithm for deriving the optimal ALPPM based on the proposed upper bound. Experimental results are presented in Section 5, followed by conclusion and future work in the next Section. Lastly, the proofs for theoretical results are given in Appendix.

2. Problem statement

In this section, we describe the system and threat model, define the privacy and utility metrics for aggregate location data, and then present the problem formulation. Table 1 gives the notations used throughout the paper.

2.1. System model

The problem we considered is illustrated by Fig. 2. In our system model, there are M users uploading their locations to a server (i.e., aggregator) who is trusted to perform aggregation. The server releases location aggregates to third-party data analysts in a privacy-preserving manner. To simplify our model, we consider users having independent location prior distributions and assume that there is no user-user or temporal correlations. If we want to blind individual locations from an untrusted server, we can aggregate the locations using cryptographic protocols (Kopp et al., 2012; Popa et al., 2011; Pyrgelis et al., 2016).

Table 1
Notations.

Symbol	Description
t, T	Timestamp (integer), time period of aggregation
l, L	Location ID (integer), total number of locations
m, M	User ID (integer), total number of users participating in the aggregation process
$\boldsymbol{A}$, $\boldsymbol{A}$	Random matrice representing the original and released location aggregates
$\boldsymbol{A}(t)$	Random vector representing the location aggregates at timestamp t
$\boldsymbol{A}_m$	Random matrix representing user m's location profile
$\boldsymbol{A}_m^{(t)}$	Random vector representing user m's location profile at timestamp t
$q(.\|.)$, $p(.,.)$	Conditional, joint probability distributions

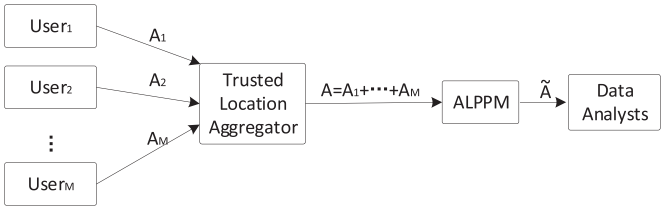


Fig. 2. Problem setting: privacy-preserving location aggregates release.

2.1.1. User's location profile

Each user's location profile is represented by a matrix $\boldsymbol{A}_m$ of size $L \times T$, where m represents user's ID, L is the total number of locations we considered (e.g., ROIs), and T is the time period that the aggregation is performed. Each column corresponding to a certain timestamp t in $\boldsymbol{A}_m$ has one element as 1 and others as 0, where 1 represents that user m visited an ROI at timestamp t and 0 denotes that she didn't.

2.1.2. User's location prior

The location prior of user m corresponding to her location profile is denoted by a matrix $\boldsymbol{P}_m$ of size $L \times T$. Each element in this matrix represents the probability of user m visiting location l at timestamp t and we have $\sum_{l=1}^{L} \boldsymbol{P}_m(l,t) = 1$. It's easy to see that user's location prior at a certain timestamp is a vector. For instance, when $L = 3$, if we assign $\boldsymbol{P}_m(l, 2)$ as $[0.1, 0.6, 0.3]^{\mathrm{T}}$, it tells us that at timestamp 2, the probabilities that user m visiting location 1, 2, 3 are 0.1, 0.6, 0.3 respectively.

2.1.3. Location aggregates

The location aggregates are represented by a matrix $\boldsymbol{A}$ of size $L \times T$. Each item $\boldsymbol{A}(l, t)$ in matrix $\boldsymbol{A}$ represents the total number of users visiting location l at timestamp t, and is calculated as $\boldsymbol{A}(l,t) = \sum_{m=1}^{M} \boldsymbol{A}_m(l,t)$. $\boldsymbol{A}(l, t)$ is an integer value between 0 and M and we have $\sum_{l=1}^{L} \boldsymbol{A}(l,t) = M$, i.e., the sum of a column equals to the total number of users who participated in the aggregation process. The matrix $\boldsymbol{A}$ can be denoted as

$$\boldsymbol{A} = \sum_{m=1}^{M} \boldsymbol{A}_m = [\boldsymbol{A}(1), ..., \boldsymbol{A}(t), ..., \boldsymbol{A}(T)], \tag{1}$$

where $\boldsymbol{A}(t)$ denotes the t-th column vector in matrix $\boldsymbol{A}$.

2.1.4. Privacy-preserving location aggregates release

In order to protect users' aggregation-based location privacy, we can perturb the aggregate location matrix while still providing certain utility before releasing it to the public. Specifically, in such a

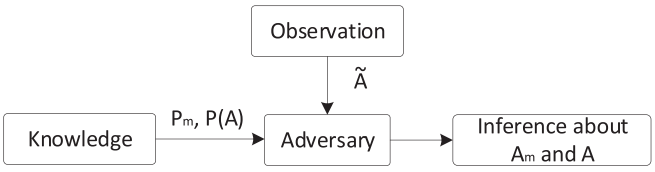


Fig. 3. Threat model.

privacy-preserving mechanism, the original aggregate location matrix $\boldsymbol{A}$ will be perturbed into its noisy version $\tilde{\boldsymbol{A}}$. Methods adopted to perturb aggregates include randomized response (Du and Zhan, 2003; Warner, 1965), noise addition (Dwork, 2008; Dwork et al., 2006), etc. To provide high data utility, the noisy aggregates should have the same properties as the original. For instance, elements in $\boldsymbol{A}$ should also be integers between 0 and M and the sum of each column should equal to M. A limitation of using noise addition is that the perturbed aggregates are not necessarily integers and could be decimals or negative or larger than M, possibly making them to be meaningless to data analysts (Wang et al., 2010). Therefore, we choose randomized response type of mechanisms to perturb the original location aggregates $\boldsymbol{A}$ since it can map each column of $\boldsymbol{A}$ into a vector with certain probability, where the vector satisfies that all the elements are integers between 0 and M and their summation equals to M, i.e., the properties of $\boldsymbol{A}$ are preserved in $\tilde{\boldsymbol{A}}$.

2.2. Threat model

To better understand the following definition of the privacy metric for location aggregates, we first model the adversary against whom the protection is provided. The figure depicting the threat model is presented in Fig. 3.

Specifically, the adversary is assumed to have full statistical knowledge of users' location priors (i.e., $\boldsymbol{P}_m$) and the location aggregates (i.e., $p(\boldsymbol{A})$). Its goal is to make inferences about a target user's location profile $\boldsymbol{A}_m$ and also the original location aggregates $\boldsymbol{A}$ after observing the perturbed location aggregates $\tilde{\boldsymbol{A}}$. Since we make no restriction on its computational capability, the adversary is theoretically capable of leveraging its own knowledge and the perturbed location aggregates $\tilde{\boldsymbol{A}}$ to perform any type of inference attack. In terms of this type of threat model, we aim to understand the fundamental information leakage (i.e., privacy leakage) on an individual user's location profile $\boldsymbol{A}_m$ as well as the original aggregates $\boldsymbol{A}$ introduced by releasing $\tilde{\boldsymbol{A}}$ from an information theoretic point of view.

2.3. Privacy and utility metrics for location aggregates

We know the leakage caused by releasing location aggregates is closely related to users' location priors and temporal correlation among locations, so they should be considered in the metrics naturally. Thus, mutual information is chosen as the privacy metric since it can capture users' location priors and the temporal correlation in a principle manner. Considering that an adversary's inference target could be an individual user's location profile $\boldsymbol{A}_m$ or the original location aggregates $\boldsymbol{A}$, we define two privacy metrics accordingly.

Definition 1Privacy Metrics for Location Aggregates.. *Given the original location aggregates $\boldsymbol{A}$, individual user's location profile $\boldsymbol{A}_m$, and the released location aggregates $\tilde{\boldsymbol{A}}$, the information leakage on an individual user's location and the original location aggregates introduced by releasing the perturbed location aggregates are defined as $I(\boldsymbol{A}_m; \tilde{\boldsymbol{A}})$ and $I(\boldsymbol{A}; \tilde{\boldsymbol{A}})$ respectively, i.e., we use mutual information as the metric to measure privacy leakage.*

In addition, we also define a utility metric to measure the utility offered by the perturbed location aggregates.

Definition 2. Utility Metric for Location Aggregates. Given the original location aggregates $\boldsymbol{A}$ and released location aggregates $\tilde{\boldsymbol{A}}$, the utility metric for location aggregates is defined as $D = D(\boldsymbol{A}; \tilde{\boldsymbol{A}}) = \sum_{t=1}^{T} D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t))$, where $D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t))$ denotes the expected distortion for the location aggregates at timestamp t, and is defined as $D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)) = \sum_{\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)} p(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)) d(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t))$, where $d(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t))$ is the Euclidean norm of vector $\boldsymbol{A}(t) - \tilde{\boldsymbol{A}}(t)$. The utility (distortion) constraint for the location aggregates at timestamp t is defined as $D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)) \leq D_t, t = 1, 2, ..., T$, which implies that the total distortion for the location aggregates $D \leq \sum_{t=1}^{T} D_t$.

2.4. Problem formulation

We first study the problem of an individual user's location privacy leakage when an adversary observes the perturbed location aggregates.

Definition 3. Individual Privacy – Aggregate Utility tradeoff of an ALPPM: Given an individual user's location profile $\boldsymbol{A}_m$, the perturbed location aggregates $\tilde{\boldsymbol{A}}$ generated by a trusted aggregator, and a utility constraint $D \leq \sum_{t=1}^{T} D_t$, the optimal ALPPM$^*_{user}$ achieves the minimum information leakage on a user's location profile subject to the utility constraint D when it is the solution of the following optimization problem:

$$ALPPM^*_{user} = \arg \min_{\substack{q(\tilde{\boldsymbol{A}}|\boldsymbol{A}): \\ \{D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)) \leq D_t\}_{t=1}^{T}}} \max_{1 \leq m \leq M} I(\boldsymbol{A}_m; \tilde{\boldsymbol{A}}), \tag{2}$$

where $I(\boldsymbol{A}_m; \tilde{\boldsymbol{A}})$ is the mutual information between $\boldsymbol{A}_m$ and $\tilde{\boldsymbol{A}}$, and the maximization is over all users. We denote all the minimum individual leakage and distortion pairs as the Individual Privacy – Aggregate Utility tradeoff $\mathcal{L}^*_{\text{user}}(D)$.

The intuition of Definition 3 is straightforward: we enumerate on all possible $q(\tilde{\boldsymbol{A}}|\boldsymbol{A})$, and for each $q(\tilde{\boldsymbol{A}}|\boldsymbol{A})$, we can get the maximal $I(\boldsymbol{A}_m; \tilde{\boldsymbol{A}})$ among all users subject to a utility constraint D ; for all the maximal $I(\boldsymbol{A}_m; \tilde{\boldsymbol{A}})$ corresponding to all $q(\tilde{\boldsymbol{A}}|\boldsymbol{A})$, we choose the $q(\tilde{\boldsymbol{A}}|\boldsymbol{A})$ that minimizes the maximum individual leakage as the optimal ALPPM (i.e., ALPPM$^*_{user}$), and consider this minimum value as the individual privacy leakage.

In addition, it's also interesting to study the privacy-utility tradeoff under the privacy metric defined as the mutual information between the original and perturbed location aggregates.

Definition 4. Aggregate Privacy – Aggregate Utility tradeoff of an ALPPM: Given the original location aggregates $\boldsymbol{A}$, the perturbed location aggregates $\tilde{\boldsymbol{A}}$ generated by a trusted aggregator, and a utility constraint $D \leq \sum_{t=1}^{T} D_t$, an ALPPM $q(\tilde{\boldsymbol{A}}|\boldsymbol{A})$ achieves the minimum information leakage on the original aggregates subject to the utility constraint D when it is the solution of the following optimization problem:

$$\mathcal{L}^*_{agg}(D) = \min_{q(\tilde{\boldsymbol{A}}|\boldsymbol{A}): \{D(\boldsymbol{A}(t), \tilde{\boldsymbol{A}}(t)) \leq D_t\}_{t=1}^{T}} I(\boldsymbol{A}; \tilde{\boldsymbol{A}}), \tag{3}$$

where $I(\boldsymbol{A}; \tilde{\boldsymbol{A}})$ is the mutual information between $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$.

Interestingly, there is an close connection between the optimization problems in Definition 3 and 4, which is proved in Section 4.

3. Preliminaries

It's easy to see that the privacy-utility tradeoffs in Definitions 3 and 4 have a close connection to the rate-distortion function in information theory. Actually, this connection has been

discussed in Zhang et al. (2019), so we will omit its analysis and only present the definition of rate-distortion function and the algorithm used for its computation, both of which are also given in Zhang et al. (2019).

Definition 5. Rate Distortion Function (Cover and Thomas, 2012). If the input of an encoder is X and the output of the corresponding decoder is $\hat{X}$, the rate distortion function $R(D)$ for a source $X \sim p(x)$ with distortion measure $d(x, \hat{x})$ is defined as

$$R(D) = \min_{\substack{p(\hat{x}|x): \\ \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D}} I(X;\hat{X}) = \min_{\substack{p(\hat{x}|x): \\ \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D}} \sum_{x,\hat{x}} p(x)p(\hat{x}|x)\log\frac{p(\hat{x}|x)}{p(\hat{x})}, \quad (4)$$

where the minimization is over all the conditional distributions $q(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ satisfies the expected distortion constraint.

3.1. Blahut-Arimoto algorithm for calculating the rate distortion function

Blahut-Arimoto algorithm (Blahut, 1972; Cover and Thomas, 2012) is an iterative algorithm which eventually converges to the optimal solution of the convex optimization problem in the rate distortion function. In detail, this algorithm works in the following procedure: it first chooses an initial probability distribution for $r(\hat{x})$ (e.g., a uniform distribution), then calculate $q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x})e^{-\lambda d(x,\hat{x})}}$ using $r(\hat{x})$. After obtaining $q(\hat{x}|x)$, it updates $r(\hat{x})$ by setting $r(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$. Then it uses $r(\hat{x})$ to update $q(\hat{x}|x)$ by setting $q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x})e^{-\lambda d(x,\hat{x})}}$. The optimal solution $q(\hat{x}|x)$ minimizing the rate distortion function is achieved by repeating the above iteration between $r(\hat{x})$ and $q(\hat{x}|x)$ until the algorithm achieves convergence.

Since Blahut-Arimoto algorithm is an efficient algorithm commonly used in information theory field to solve optimization problems that can be formulated as a rate-distortion problem, we modify it to suit our problem in this paper.

In this work, we also study the privacy-utility tradeoff in the case where all users' location priors are the same. For example, when only a global prior is known for all the users instead of individual users' priors, which could be obtained from historically collected data such as population census or surveys, the probability distribution of each column in $\boldsymbol{A}$ is actually a multinomial distribution. Therefore, we briefly describe its definition below.

Definition 6. Multinomial Distribution (Jaynes, 2003). The random vector RV $\vec{\boldsymbol{X}} = (X_1, ..., X_m)$ has a multinomial distribution with parameters $N \in \{1, 2, ...\}$ and $\theta \in \mathbb{R}^n$ for all i and $\sum_{i=1}^m \theta_i = 1$ if

$$p_{\vec{\boldsymbol{X}}}(x_1, ..., x_m) = \begin{cases} \binom{N}{x_1,...,x_m}\theta_1^{x_1}\cdots\theta_m^{x_m}, & \text{if } x_1, ...,_m \\ \text{are non-negative} \\ \text{integers that sum to N} \\ 0, textotherwise \end{cases}$$

where $\binom{N}{x_1,...,x_m} = \frac{N!}{x_1!\cdots x_m!}$ is the multinomial coefficient.

4. Main results and algorithm

4.1. Practical challenge

We notice that there is a challenge in finding the optimal privacy-utility tradeoff for location aggregates, i.e., the exponential complexity caused by directly using Blahut-Arimoto algorithm on the optimization problem in Definition 4. This is because we have to characterize $q(\boldsymbol{A}|\boldsymbol{A})$ for all $(\boldsymbol{A}, \boldsymbol{A}) \in \mathcal{A} \times \mathcal{A}$, meaning that the optimization problem has to be solved over $|\mathcal{A}||\mathcal{A}|$ variables to find the optimal solution $q(\boldsymbol{A}|\boldsymbol{A})$. For instance, if we assume that there are N realizations of each column, the total number of variables in the optimization problem will be N^{2T}. Either the increase of N or T will make the number of variables increase exponentially.

To avoid the complexity caused by T, we make an assumption that no temporal correlations exist among all the timestamps in the aggregation process. This scenario is reasonable when the collected location traces are sporadic (Andrés et al., 2013; Oya et al., 2017; Shokri et al., 2012). Next, we will show how to derive an upper bound on $\mathcal{L}^*_{agg}(D)$ based on this assumption and then present an algorithm to solve the optimization problem in the upper bound.

4.2. Main results

The following theorem shows the upper and lower bounds on the Aggregate Privacy – Aggregate Utility tradeoff.

Theorem 1. *When there is no temporal correlation in the location aggregates, the main results for individual and aggregate privacy-utility tradeoff are:*

$$\mathcal{L}^*_{\text{user}}(D) \le \mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg}) \le \mathcal{L}^*_{\text{agg}}(D) \le \mathcal{L}^{\text{upper}}_{\text{agg}}(D), \quad (5)$$

where

$$\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg}) = \max_{1\le m\le M} I(\boldsymbol{A}_m; \boldsymbol{A})|_{q(\boldsymbol{A}|\boldsymbol{A})=\text{ALPPM}^*_{agg}} \quad (6)$$

and

$$\mathcal{L}^{\text{upper}}_{\text{agg}}(D) = \sum_{i=1}^{T} \min_{\substack{q(\boldsymbol{A}(t)|\boldsymbol{A}(t)): \\ D(\boldsymbol{A}(t),\boldsymbol{A}(t))\le D_t}} I(\boldsymbol{A}(t); \boldsymbol{A}(t)). \quad (7)$$

The proof of Theorem 1 is given in Appendix.

Since generating ALPPMs according to $\mathcal{L}^*_{user}(D)$ and $\mathcal{L}^*_{agg}(D)$ incurs exponential complexity, we will leverage $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$ to generate the optimal aggregation-based location privacy preserving mechanism ALPPM^*_{agg}, which can provide the privacy guarantee that the actual leakage for location aggregates is upper bounded by $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$. Moreover, Theorem 1 also shows $\mathcal{L}^*_{user}(D) \le \mathcal{L}^*_{agg}(D)$, i.e., the leakage on an individual user's location is upper bounded by the leakage on the original aggregates introduced by the perturbed aggregates. More importantly, we have proved an tighter upper bound on $\mathcal{L}^*_{user}(D)$, which is $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$, i.e., the maximal individual user's leakage under ALPPM^*_{agg}.

However, it's not clear how much actual leakage occurs on the original location aggregates when releasing the perturbed aggregates according to ALPPM^*_{agg}. Accordingly, we present the following definition to calculate the amount of actual leakage.

Definition 7. Actual Privacy Leakage of Location Aggregates Without Temporal Correlation. When there is no temporal correlation among location aggregates, for a certain time period 1, 2, ..., T, the actual privacy leakage of an ALPPM is defined as

$$\mathcal{L}_{\text{actual}}(\text{ALPPM}) = \sum_{i=1}^{T} I(\boldsymbol{A}(t); \boldsymbol{A}(t)), \quad (8)$$

where the ALPPM can be generated based on any type of approaches.

Interestingly, the following corollary proves that the actual leakage of ALPPM^*_{agg} generated according to $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$ equals to $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$.

Corollary 1. *The actual privacy leakage of* ALPPM^*_{agg} *evaluated by* $\mathcal{L}_{\text{actual}}(\text{ALPPM})$ *equals to* $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$, *i.e.,*

$$\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg}) = \mathcal{L}^{\text{upper}}_{\text{agg}}(D). \quad (9)$$

The proof of Corollary 1 is trivial. It's readily seen that $\sum_{i=1}^{T} I(\boldsymbol{A}(t);\boldsymbol{A}(t))$ equals to the summation of the objective functions in $\mathcal{L}_{\text{agg}}^{\text{upper}}(D)$. We know ALPPM_{agg}^{*} is generated according to $\mathcal{L}_{\text{agg}}^{\text{upper}}(D)$, thus we have $\mathcal{L}_{\text{actual}}(\text{ALPPM}_{agg}^{*}) = \mathcal{L}_{\text{agg}}^{\text{upper}}(D)$.

4.3. Algorithm

In this part, we propose an algorithm based on Blahut-Arimoto algorithm (Blahut, 1972; Cover and Thomas, 2012) to obtain the optimal aggregation-based location privacy-preserving mechanism ALPPM_{agg}^{*} according to the upper bound $\mathcal{L}_{\text{agg}}^{\text{upper}}(D)$, which is presented in Algorithm 1 . Essentially, we leverage its basic idea to implement an iterative algorithm which eventually converges to the optimal solution of minimizing the mutual information between two vectors. Specifically, in Algorithm 1, λ is the Lagrange multiplier used in solving the optimization problem and represents how much we favor information leakage versus distortion (smaller λ means more distortion), and δ is the threshold for the proposed algorithm to achieve convergence. Next, we describe how to obtain the other two inputs $\boldsymbol{a}(t)$ and $p(\boldsymbol{a}(t))$.

Algorithm 1: Generating ALPPM_t at timestamp t

Input: $p(\boldsymbol{a}(t))$: probability distribution of $\boldsymbol{a}(t)$, λ: Lagrange multiplier, $D(\boldsymbol{a}(t),\boldsymbol{a}(t))$: distortion matrix, δ: threshold for convergence

Output: $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$: ALPPM at timestamp t, I_t^*: minimum leakage at timestamp t, D_t: distortion corresponding to I_t^*

1: Initialize $r_0(\boldsymbol{a}(t))$ as a uniform distribution
2: Calculate $q_0(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ using $r_0(\boldsymbol{a}(t))$ by $q(\boldsymbol{a}(t)|\boldsymbol{a}(t)) = \frac{r_0(\boldsymbol{a}(t))e^{-\lambda d(\boldsymbol{a}(t),\boldsymbol{a}(t))}}{\sum_{\boldsymbol{a}(t)} r_0(\boldsymbol{a}(t))e^{-\lambda d(\boldsymbol{a}(t),\boldsymbol{a}(t))}}$
3: Calculate I_t^0 using $r_0(\boldsymbol{a}(t))$, $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$, and $p(\boldsymbol{a}(t))$ by $I_t^0 = \sum_{\boldsymbol{a}(t),\boldsymbol{a}(t)} p(\boldsymbol{a}(t))q(\boldsymbol{a}(t)|\boldsymbol{a}(t)) \log \frac{q(\boldsymbol{a}(t)|\boldsymbol{a}(t))}{r_0(\boldsymbol{a}(t))}$
4: Calculate $r(\boldsymbol{a}(t))$ using $q_0(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ by $r(\boldsymbol{a}(t)) = \sum_{\boldsymbol{a}(t)} p(\boldsymbol{a}(t))q_0(\boldsymbol{a}(t)|\boldsymbol{a}(t))$
5: **while** true **do**
6: Calculate $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ using $r(\boldsymbol{a}(t))$ by $q(\boldsymbol{a}(t)|\boldsymbol{a}(t)) = \frac{r(\boldsymbol{a}(t))e^{-\lambda d(\boldsymbol{a}(t),\boldsymbol{a}(t))}}{\sum_{\boldsymbol{a}(t)} r(\boldsymbol{a}(t))e^{-\lambda d(\boldsymbol{a}(t),\boldsymbol{a}(t))}}$
7: Calculate I_t using $r(\boldsymbol{a}(t))$, $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$, and $p(\boldsymbol{a}(t))$ by $I_t = \sum_{\boldsymbol{a}(t),\boldsymbol{a}(t)} p(\boldsymbol{a}(t))q(\boldsymbol{a}(t)|\boldsymbol{a}(t)) \log \frac{q(\boldsymbol{a}(t)|\boldsymbol{a}(t))}{r(\boldsymbol{a}(t))}$
8: **if** $(I_t^0 - I_t \leqslant \delta)$ **then**
9: $I_t^* \leftarrow I_t$
10: Calculate $D_t = \sum_{\boldsymbol{a}(t),\boldsymbol{a}(t)} p(\boldsymbol{a}(t))q(\boldsymbol{a}(t)|\boldsymbol{a}(t))d(\boldsymbol{a}(t),\boldsymbol{a}(t))$
11: **return** $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$, I_t^*, D_t,
12: **else**
13: $I_t^0 \leftarrow I_t$
14: Calculate $r(\boldsymbol{a}(t))$ using $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ by $r(\boldsymbol{a}(t)) = \sum_{\boldsymbol{a}(t)} p(\boldsymbol{a}(t))q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$
15: **end if**
16: **end while**

If we assume the number of users is M and the number of locations is L, $p(\boldsymbol{a}(t))$ is calculated by function $EC(M, L)$ and $CPoC(\boldsymbol{V})$[1], where $\boldsymbol{a}(t)$ is a random vector representing all users' visits in L locations at timestamp t, and $p(\boldsymbol{a}(t))$ is its probability distribution. In detail, function $EC(M, L)$ is designed to characterize $\boldsymbol{a}(t)$ by enumerating all the cases of distributing M users into L locations, and function $CPoC(\boldsymbol{V})$ is used to calculate $p(\boldsymbol{a}(t))$ by computing the probability of each case. For example, when $M = 4, L = 2$, we have all possible outputs of $EC(M, L)$ as $C = ((0,4),(1,3),(2,2),(3,1),(4,0))$. When all users' location priors are the same, $\boldsymbol{a}(t)$ is actually a multinomial random vector. In this case, function $CPoC(\boldsymbol{V})$ is defined according to the definition of multinomial distribution presented in Section 3, so the output of $CPoC(\boldsymbol{V})$ is $p(\boldsymbol{V}) = \frac{M!}{v_1!\cdots v_L!}\theta_1^{v_1}\cdots\theta_L^{v_L}$, where θ_l is the probability of a user visiting location l and $\sum_{l=1}^{L}\theta_l = 1$. When users' location priors are different, $p(\boldsymbol{a}(t))$ can be calculated according to the law of total probability. Clearly, the release of the aggregate location matrix $\boldsymbol{A}$ is achieved by releasing $\boldsymbol{A}(t)$ at each timestamp t according to ALPPM_t generated from Algorithm 1 at all timestampss.

[1] EC and CPoC are short for EnumerateCombinations and CalcualteProbabilityOfCombinations respectively.

Table 2
M and its Corresponding Maximum L.

M, L_{max}	M, L_{max}	M, L_{max}	M, L_{max}	M, L_{max}
1,53	8,17	15,14	22,12	29,11
2,33	9,16	16,13	23,12	30,11
3,27	10,16	17,13	24,12	31,11
4,23	11,15	18,13	25,12	32,11
5,21	12,15	19,13	26,11	33,11
6,19	13,14	20,12	27,11	34,11
7,18	14,14	21,12	28,11	35,11

To show the complexity of Algorithm 1, we present an expression for the computation complexity of one iteration in this algorithm, which is similar to the analysis in Zhang et al. (2019). In each iteration, the computation complexity is dominated by the calculation of $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ and $r(\boldsymbol{a}(t))$. (1): According to the equation given in step 2 in Algorithm 1, it's easy to see that for each $\boldsymbol{a}(t)$, we need to do $|\mathcal{A}(t)|$ multiplications for a specific $\boldsymbol{a}(t)$ in the denominator, and then use this denominator for every other $\boldsymbol{a}(t)$, so the computation needs $O(|\mathcal{A}(t)|)$ operations. Considering of all $\boldsymbol{a}(t)$, the complexity of calculating $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$ is $O(|\mathcal{A}(t)||\mathcal{A}(\boldsymbol{t})|)$. (2) : According to the equation given in step 4, we need to do $|\mathcal{A}(\boldsymbol{t})|$ multiplications for a specific $\boldsymbol{a}(t)$. Considering of the calculation for all $\boldsymbol{a}(t)$, the complexity of computing $r(\boldsymbol{a}(t))$ is also $O(|\mathcal{A}(t)||\mathcal{A}(\boldsymbol{t})|)$. As a result, each iteration in Algorithm 1 requires about $O(|\mathcal{A}(t)||\mathcal{A}(\boldsymbol{t})|)$ computations. If we assume there are N realizations of each column, $|\mathcal{A}(t)||\mathcal{A}(\boldsymbol{t})|$ equals to N^2, which is much less than the computational complexity in the optimization problem in Definition 4 (i.e., N^{2T}). The issue that $|\mathcal{A}(t)|$ and $|\mathcal{A}(\boldsymbol{t})|$ increase exponentially with the increase of the number of locations or users will be addressed in future work.

5. Experimental evaluation

In this section, we evaluate the actual privacy leakage of our proposed ALPPM and compare it with a differentially private mechanism over both synthetic and real-world datasets. The compared one is an output perturbation mechanism called Simple Counting Mechanism (SCM), where a trust sever adds noise to the aggregate before releasing it to the public (Chan et al., 2011). Specifically, SCM samples random values from a Laplace distribution $\text{LAP}(1/\epsilon)$ since the sensitivity of counting queries is 1. All evaluation were conducted on a desktop with 2.40 GHz Intel i5 CPU and 8GB memory.

5.1. Simulation setup

Considering the limitation of a computer's memory, there exits a maximal number of operations that a computer can hold in its memory and perform calculations with. In the following, we make a list of the number of users (M) and the corresponding maximal number of locations (L_{max}), which are possible values of the inputs in Algorithm 1. As shown in Table 2, the maximal value of M that

Algorithm 1 can process is 35. Designing algorithms which support larger inputs M and L is considered as future work.

In this simulation, we generate 4 synthetic datasets corresponding to the cases when there are 5, 10, 15 and 20 users participating in aggregation on 3 different locations, so as to explore the impact of the number of users on the privacy-utility tradeoff. Also, in order to see how users' location priors affect the privacy-utility tradeoff, we conduct the evaluation of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ under both homogeneous and different user priors on the synthetic datasets. To present the figures for experimental results in a clear way, we use the letters ALPPM as an abbreviation for $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and SCM as an abbreviation for $\mathcal{L}_{\text{actual}}(\text{SCM})$ in all the following figures.

5.2. Evaluation on synthetic datasets

As dominant parameters used to generate aggregate location dataset, the impacts of the number of users M and users' location priors need to be carefully analyzed. More importantly, it's also interesting to study how the skewness of users' location priors affects the privacy-utility tradeoff. Knowing from Corollary 1 that the actual leakage of the ALPPM generated according to $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$ equals to $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$, i.e., $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg}) = \mathcal{L}^{\text{upper}}_{\text{agg}}(D)$, we only present the results of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and compare it to $\mathcal{L}_{\text{actual}}(\text{SCM})$ under the same distortion. Specifically, we analyze the privacy-utility tradeoffs when users' priors are the same and different separately.

5.2.1. Evaluation of $\mathcal{L}_{actual}(ALPPM^*_{agg})$ and $\mathcal{L}_{actual}(SCM)$ under homogeneous user priors

In this part, we present the simulation results of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ when users have the same location priors. The evaluation was done on four different types of priors at a single timestamp t, which are $p_1 = [1/3, 1/3, 1/3]^{\text{T}}$, $p_2 = [0.25, 0.5, 0.25]^{\text{T}}$, $p_3 = [0.1, 0.1, 0.8]^{\text{T}}$ and $p_4 = [0.8, 0.1, 0.1]^{\text{T}}$ respectively, to see the impact of skewness of location prior distributions. Specifically, p_1 represents the case where users visit each location equally likely, p_3 and p_4 are designed in a way that users have much higher chances visiting certain locations than others, while p_2 shows users have slightly higher probability visiting certain locations than other locations. It's easy to see that p_3 and p_4 have the largest skewness while p_1 has the least. The case where location priors have large skewness actually has lots of applications, since it can represent tourist attractions, sport stadiums, shopping malls, or schools. The skewness of location prior is worth of analyzing, because a skewed prior means there are popular locations and the analysis on popular locations helps to provide insight for choosing optimal geographic placement for retail stores or advertisements.

Firstly, the probability distributions of aggregate location data are presented in Fig. 4 to show how they are influenced by the skewness of location priors. To ease presentation, we only show the probability distributions when $L = 3, M = 5$ and users' priors are p_1, p_2, p_3 and p_4. It's easy to see that probability distribution of different categories (i.e., possible combinations of 5 users locating at 3 locations at timestamp t) have different levels of skewness. Specifically, Fig. 4a shows categories [1 2 2], [2 1 2], and [2 2 1] have the highest probability, while [0 0 5], [0 5 0], and [5 0 0] have the lowest probability. This is because users' priors are uniformly distributed, meaning that they visit each location equally likely. In Fig. 4b, [1 3 1] has the highest probability in all categories since the second location is more popular than the other two in p_2. Obviously, the probability distribution in Fig. 4b is more skewed than the one in Fig. 4a. The distributions in Fig. 4c and 4 d have the same degree of skewness as the location popularity in p_3 and p_4 are the same, and they are more skewed than the one in Fig. 4b. For example, the highest probabilities in Fig. 4c and 4 d happen on category [0 0 5] and [5 0 0] respectively, because the third location in p_3 and the first location in p_4 are the most popular ones. Therefore, we can conclude that the level of skewness is related to the popularities of certain locations in users' priors. In other words, when certain locations in users' priors have much higher probabilities than the others, the probability distribution of aggregate location will have larger skewness.

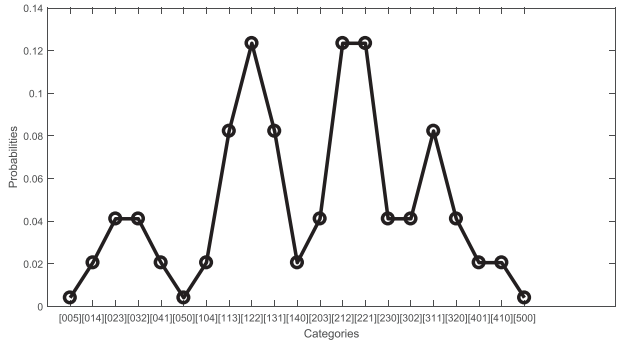
Next, we show the evaluation of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on p_1, p_2, p_3 and p_4 when the total number of users changes in Fig. 5. We start with describing the details about how to derive the curves shown in the figures below. Each point on the curve is corresponding to a fixed λ, i.e., the Lagrange multiplier, which is one of the inputs in Algorithm 1. For a fixed λ, given users' priors, we can obtain an information leakage-distortion pair by Algorithm 1 and save the output distortion D. When changing to different λs, we can smoothly draw the information leakage-distortion curve of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$. Now we explain how to derive the SCM under the distortion saved earlier. For every λ, we increase ϵ (i.e., the privacy parameter in differential privacy) from 0.001 to 100 and use the same users' priors to derive the SCM under the same distortion as D. According to the conditional probability distribution obtained from SCM, we can easily calculate its actual aggregate leakage by Definition 7. By enumerating all the λs, we draw the curves of $\mathcal{L}_{\text{actual}}(\text{SCM})$. In addtion, the threshold for convergence in Algorithm 1 is set as 0.001 and its reason will be explained in Section 5.3.

Overall, $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ is always lower than $\mathcal{L}_{\text{actual}}(\text{SCM})$ in all cases, which means the actual information leakage of the proposed ALPPM is less than the actual leakage of SCM under the same distortion. Regarding the privacy-utility tradeoff curves of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$, it's easy to see that the curves corresponding to p_3 and p_4 coincide, since p_3 and p_4 have the same degree of skewness, meaning that the probabilities used to calculate the actual information leakage are the same. Another important insight is that given the same distortion, the least leakage occurs on p_3 and p_4, while the most leakage happens on p_1. This is because a skewed prior itself has already revealed much information, while our mechanism minimizes the additional leakage after releasing the perturbed aggregates. As a result, the privacy-utility tradeoff curves corresponding to the priors as p_3 and p_4 are the lowest compared to the cases when priors are p_1 and p_2. The results provide us a takeaway that the proposed mechanism can guarantee that the more skewed users' priors are, the less leakage will occur after releasing perturbed location aggregates. In other words, location aggregates are better protected in the case of skewed priors than uniform priors. Actually, users tend to have skewed priors since real-world location distributions are heterogeneous.

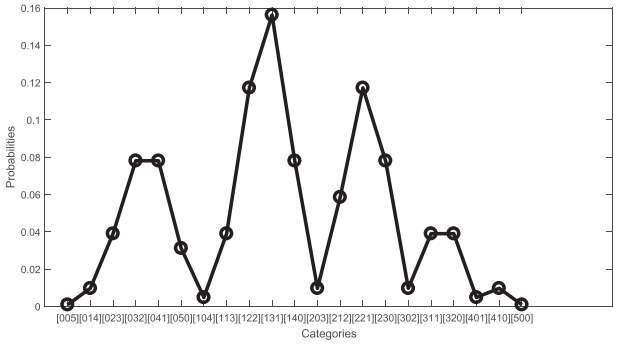
Interestingly, $\mathcal{L}_{\text{actual}}(\text{SCM})$ shows the similar patterns as $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$, but we will omit its analysis due to space consideration. We want to highlight that even though directly computing $\mathcal{L}^*_{\text{user}}(D)$ is challenging due to exponential complexity, releasing location aggregates according to ALPPM^*_{agg} can still provide the privacy guarantee that the maximum leakage of individual users is upper bounded by $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$. Lastly, we also check the conditional probabilities of outputting popular and unpopular locations given by $q(\boldsymbol{A}(t)|\boldsymbol{A}(t))$, and they have shown that popular locations are less perturbed while unpopular ones are perturbed more, which means unpopular locations are better protected than popular locations.

5.2.2. Evaluation of $\mathcal{L}_{actual}(ALPPM^*_{agg})$ and $\mathcal{L}_{actual}(SCM)$ when users' priors are different

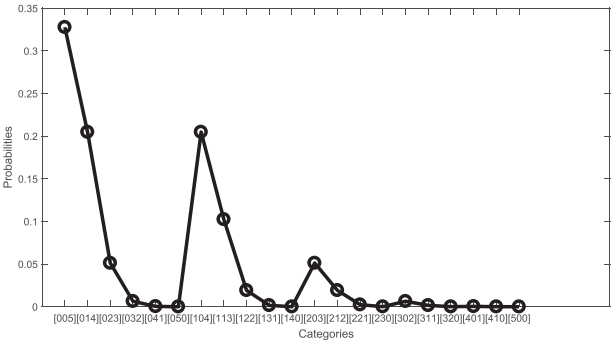
In real-world scenarios, users' priors tend to be different, so we also evaluate $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ in this case. We set $L = 3, M = 5$ to ease presentation. To explore how the skewness



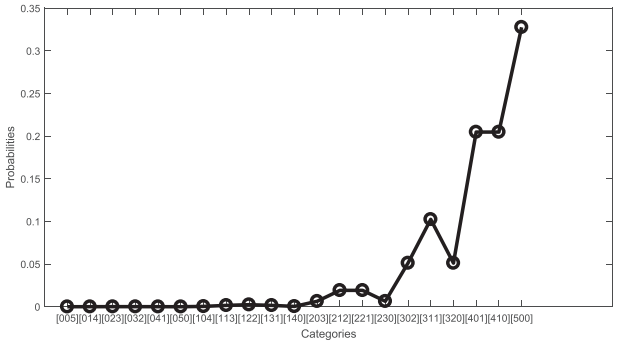
(a) Users' prior is p_1.



(b) Users' prior is p_2.

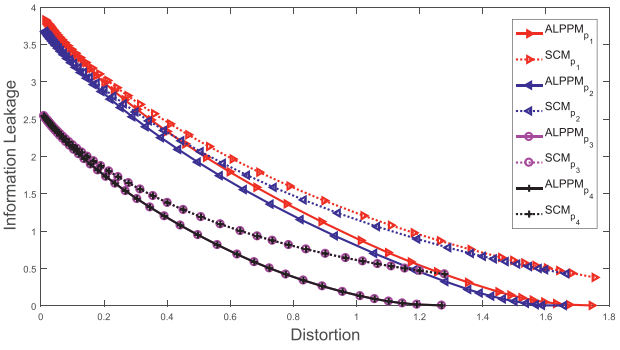


(c) Users' prior is p_3.

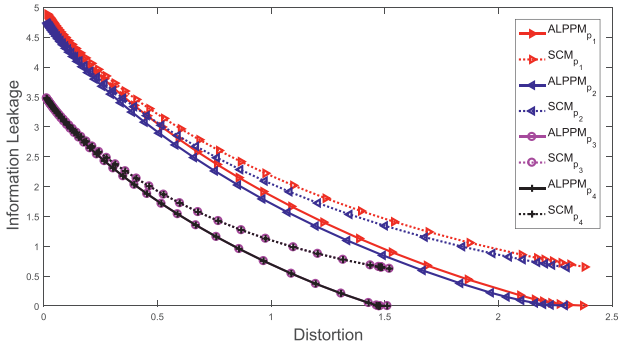


(d) Users' prior is p_4.

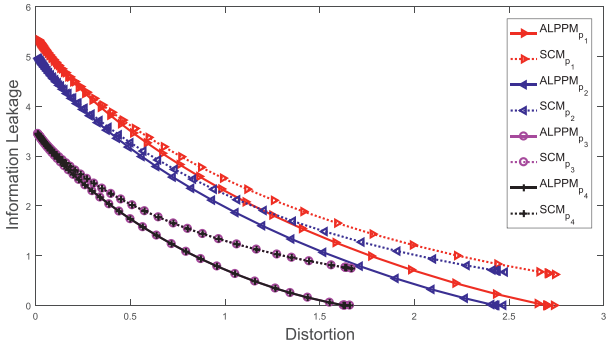
Fig. 4. PDF of aggregated location data on 4 different users' priors: $p_1 = [1/3, 1/3, 1/3]^{\mathsf{T}}$, $p_2 = [0.25, 0.5, 0.25]^{\mathsf{T}}$, $p_3 = [0.1, 0.1, 0.8]^{\mathsf{T}}$ and $p_4 = [0.8, 0.1, 0.1]^{\mathsf{T}}$ when $L = 3, M = 5$.



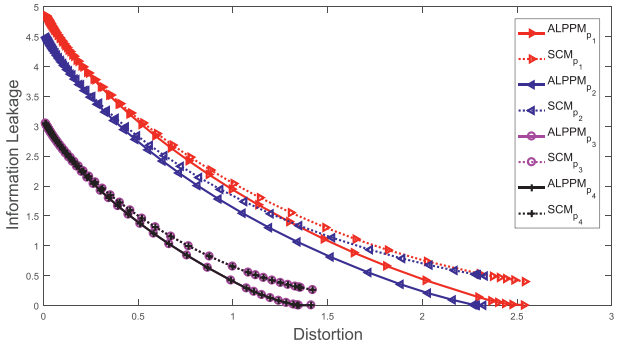
(a) M=5,L=3.



(b) M=10,L=3.

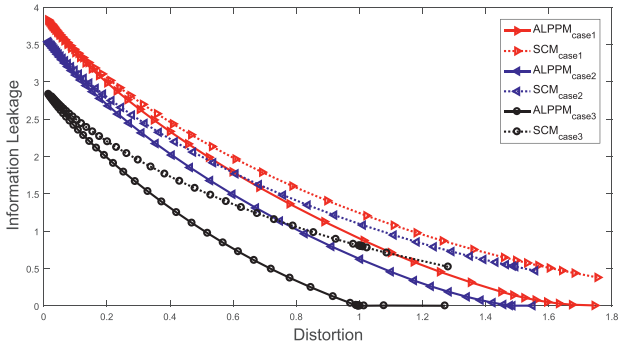


(c) M=15,L=3.

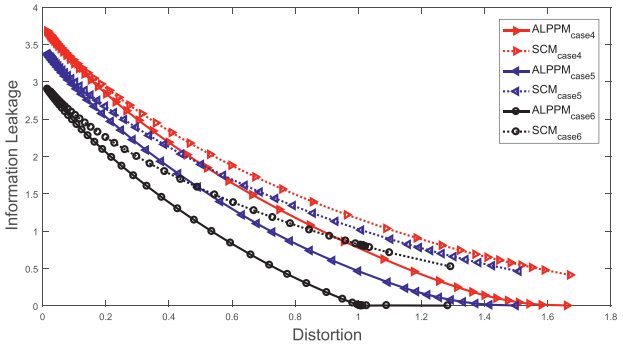


(d) M=20,L=3.

Fig. 5. Evaluation of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on 4 types of user priors: $p_1 = [1/3, 1/3, 1/3]^{\mathsf{T}}$, $p_2 = [0.25, 0.5, 0.25]^{\mathsf{T}}$, $p_3 = [0.1, 0.1, 0.8]^{\mathsf{T}}$ and $p_4 = [0.8, 0.1, 0.1]^{\mathsf{T}}$.

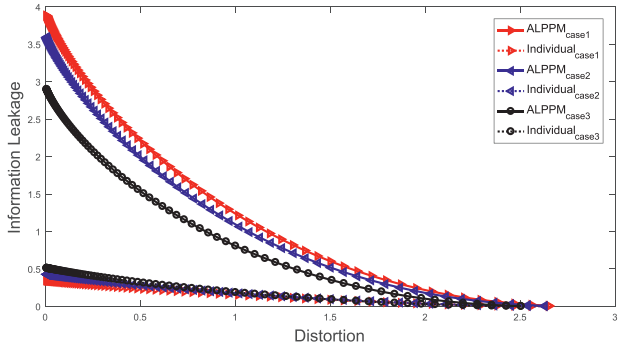


(a) Comparison among case 1, 2, 3.

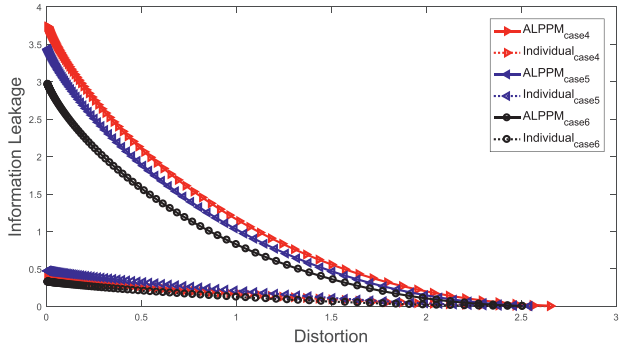


(b) Comparison among case 4, 5, 6.

Fig. 6. Evaluation of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ when users' priors are different and $L = 3, M = 5$.



(a) Comparison among case 1, 2, 3.



(b) Comparison among case 4, 5, 6.

Fig. 7. Comparison between $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$ (denoted by Individual) and $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ (denoted by ALPPM) when $L = 3, M = 5$.

of users' location priors affects the privacy leakage, we consider six cases, each of which represents one possible group of users participating in an aggregation process. Each case is denoted as a matrix named C_i with size 5×3, where every row corresponds to an individual user's location prior, and the six matrices are shown below. C_1 represents an extreme case that all users visit each location equally likely, i.e., their location priors have the lowest skewness. C_2 represents a scenario where every user's prior is randomly generated. C_3 is another case where each user has certain location with higher popularity than the others. The level of skewness of C_2 is between C_1 and C_3. We compare the results of C_1, C_2 and C_3 in Fig. 6a. In addition, the evaluations on C_4, C_5 and C_6 are designed to help us to better understand how the number of users who have higher chances to visit certain locations affect the privacy leakage, and the results are presented in Fig. 6b.

$$C_1 = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad C_2 = \begin{bmatrix} 0.421 & 0.129 & 0.450 \\ 0.415 & 0.219 & 0.366 \\ 0.188 & 0.777 & 0.035 \\ 0.659 & 0.110 & 0.231 \\ 0.449 & 0.379 & 0.172 \end{bmatrix}$$

$$C_3 = \begin{bmatrix} 0.06 & 0.9 & 0.04 \\ 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.05 & 0.85 \\ 0.75 & 0.15 & 0.1 \end{bmatrix} \quad C_4 = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$C_5 = \begin{bmatrix} 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad C_6 = \begin{bmatrix} 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$$

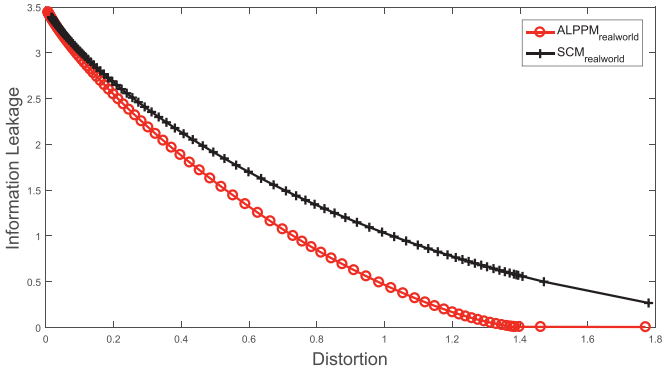
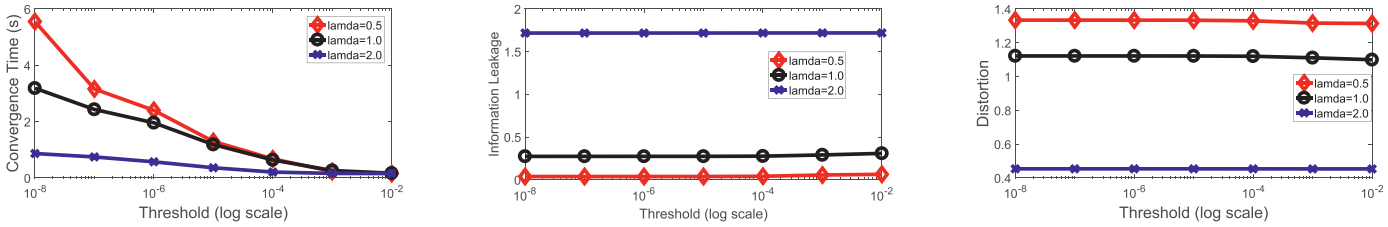


Fig. 8. Evaluation of $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on a real-world dataset.

It's clear from Fig. 6 that $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ is always lower than $\mathcal{L}_{\text{actual}}(\text{SCM})$ when users' location priors are different. Specifically, in the case of users visiting each location equally likely, the privacy leakage is the largest. It also shows that the more users who have higher probabilities visiting certain locations instead of visiting each location equally likely, the less privacy leakage will occur after releasing perturbed location aggregates according to ALPPM^*_{agg}. This is because when a location prior is highly skewed, the prior itself has already revealed quite much information. For instance, compared with C_3, an adversary can infer less information about user's locations by only knowing the priors in C_1 and C_2. In other words, when a user's location prior is a uniform distribution, it's almost impossible for an adversary to guess which ROI she visited by making an inference only based on her prior (i.e., without the perturbed aggregates). Once the adversary observes the perturbed



(a) Impact on runtime for convergence. (b) Impact on information leakage. (c) Impact on distortion.

Fig. 9. Impact of threshold.

aggregates, he will be able to learn more information than the case when users' priors are more skewed. Therefore, when publishing aggregate location matrix, the proposed mechanism ensures that the more skewed users' priors are, the better protection it can provide to the original location aggregates. Moreover, its advantage over SCM in terms of privacy-utility tradeoff becomes even greater when there are more users with highly skewed priors.

In addition, we also calculate the maximal individual leakage under ALPPM^*_{agg}, i.e., $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$, and compare it with $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$. Specifically, if we denote the optimal mechanism ALPPM^*_{agg} at timestamp t as $q(\boldsymbol{a}(t)|\boldsymbol{a}(t))$, we can calculate $q(\boldsymbol{a}(t)|\boldsymbol{a}_m(t))$ according to the law of total probability $q(\boldsymbol{a}(t)|\boldsymbol{a}_m(t)) = \sum_{a(t)} q(\boldsymbol{a}(t)|\boldsymbol{a}(t))q(\boldsymbol{a}(t)|\boldsymbol{a}_m(t))$. Then we can compute each user's information leakage as $I(\boldsymbol{A}_m(t); \boldsymbol{A}(t))$ and take the maximum as the individual leakage. By calculating the maximal individual user's leakage for each λ, we can draw the information leakage and distortion curves for $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$ and compare them with $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ in Fig. 7. Results show that $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$ is always less than $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ in all cases, which is also proved to be true in Theorem 1. Even though we don't know the exact privacy-utility tradeoff for $\mathcal{L}^*_{\text{user}}(D)$, releasing perturbed location aggregates according to ALPPM^*_{agg} can still ensure that $\mathcal{L}^*_{\text{user}}(D)$ is no larger than $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$, i.e., the privacy leakage of each user participating in the aggregation is always less than $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$.

5.3. Evaluation on real-world dataset

In this part, we evaluate $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on a real-world dataset Gowalla (Cho et al., 2011), which is a location-based social networking website where users share their locations by checking-in. Since computing $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on the entire dataset is impracticable due to the limitation of a computer's memory, we randomly select 5 users out of the entire dataset to train their personal location probability distributions (i.e., priors) on 3 locations, and then calculate $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ and $\mathcal{L}_{\text{actual}}(\text{SCM})$ on the preprocessed dataset generated by those 5 users. In particular, we first round these users' location IDs to 2 significant digits (counted from the left-most digit), then we count their occurrences on each location, and finally we choose the top 3 most frequent location IDs to obtain the preprocessed dataset. We set λ as the same range as in the synthetic dataset and set the threshold as 0.001.

Results presented in Fig. 8 show that $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ is always less than $\mathcal{L}_{\text{actual}}(\text{SCM})$ in the real-world dataset, similar to the results on synthetic datasets. These results support the idea that individual users who participate in an aggregation process can be guaranteed that their privacy leakage will be no larger than $\mathcal{L}_{\text{actual}}(\text{ALPPM}^*_{agg})$ by releasing location aggregates according to $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$. Therefore, they will be willing to contribute to location aggregation without privacy concerns, which is important for generating valuable aggregate location datasets for statistical analysis while still protecting users' location privacy.

As we can see in this section, assigning a value to the threshold is an essential step. In order to clearly explain how to choose a proper value for the threshold and its physical significance, we present the following results in Fig. 9, which show how different thresholds affect the time required for Algorithm 1 to achieve convergence, the information leakage and the distortion. It's easy to see from Fig. 9 that the time needed for convergence for Algorithm 1 decreases noticeably with the increase of threshold, while the information leakage and distortion only change slightly. These results provide guidance on how to assign proper values to the threshold. Now it's easy to understand the reason for the threshold to be set as 0.001 in all simulations is that we can guarantee the runtime is shorter compared with a smaller threshold while the corresponding information leakage only has a slight increase.

6. Conclusion and future work

We have proposed privacy metrics to measure aggregation-based location privacy independent of any specific attack based on an information-theoretic approach, and formulated the problem to obtain the optimal ALPPM used to release location aggregates while achieving the minimum information leakage given a utility constraint. To address the computation challenge occurred when computing the optimal ALPPM, we obtain an upper bound on the privacy-utility tradeoff when there is no temporal correlation in the aggregation process. Experiments have shown the actual leakage of the proposed ALLPM is less than the leakage of a differentially private mechanism under the same distortion and this advantage is greater when users visit certain locations with higher probabilities. In addition, the proposed privacy metrics can also be used as standard measures to evaluate and compare other privacy-preserving mechanisms for aggregate statistics, which is useful in many real-world applications. Our future work includes taking temporal and user correlations into account and improving algorithms for calculations on larger M and L.

Declaration of Competing Interest

None.

Acknowledgement

This work was supported by NSFC 61932015 and in part by U.S. NSF under Grants CNS-1731164. The work of Ravi Tandon was supported in part by NSF grants CAREER 1651492, CNS 1715947 and the 2018 Keysight Early Career Professor Award. Hui Li and part of Wenjing Zhang's work were supported by NSFC 6173202, SHAANXI Innovation team project 2018TD-007, 111 project B16037. Part of this work was done when Wenjing Zhang visited the Department of Electrical and Computer Engineering at The University of Arizona.

Appendix A

A1. Proof of Theorem 1

A1.1. $\mathcal{L}^*_{agg}(D) \le \mathcal{L}^{upper}_{agg}(D)$

We start with proving the connection between the objective function in $\mathcal{L}^*_{\text{agg}}(D)$ and the one in $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$,

$$
\begin{aligned}
&I(\boldsymbol{A};\boldsymbol{A}) \\
&\quad = I(\boldsymbol{A}(1),...,\boldsymbol{A}(T);\boldsymbol{A}(1),...,\boldsymbol{A}(T)) \\
&(a) = \sum_{i=1}^{T} I(\boldsymbol{A}(i);\boldsymbol{A}(1),...,\boldsymbol{A}(T)|\boldsymbol{A}(1),...,\boldsymbol{A}(i-1)) \qquad (11)\\
&(b) = \sum_{i=1}^{T}\sum_{j=1}^{T} I(\boldsymbol{A}(i);\boldsymbol{A}(j)|\boldsymbol{A}(1),...,\boldsymbol{A}(i-1), \qquad (12)\\
&\boldsymbol{A}(1),...,\boldsymbol{A}(j-1)) \qquad (13)\\
&= \sum_{i=j=1}^{T} I(\boldsymbol{A}(i);\boldsymbol{A}(j)|\boldsymbol{A}(1),...,\boldsymbol{A}(i-1), \\
&\quad \boldsymbol{A}(1),...,\boldsymbol{A}(j-1)) \\
&\quad + \sum_{i=1}^{T}\sum_{j=1,j\neq i}^{T} I(\boldsymbol{A}(i);\boldsymbol{A}(j)|\boldsymbol{A}(1),...,\boldsymbol{A}(i-1) \\
&\quad \boldsymbol{A}(1),...,\boldsymbol{A}(j-1)) \qquad (14)\\
&(c) = \sum_{i=1}^{T} I(\boldsymbol{A}(i);\boldsymbol{A}(i)), \qquad (15)
\end{aligned}
$$

where (a) and (b) follows from the chain rule of mutual information, (c) follows from the fact that $\boldsymbol{A}(i)$ is independent of $\boldsymbol{A}(i)$ when $i \neq j$, since we have the assumption that there is no temporal correlation among all the timestamps in the aggregation process.

The above proof shows the objective function in $\mathcal{L}^*_{\text{agg}}(D)$ equals to the summation of the objective functions in $\mathcal{L}^{\text{upper}}_{\text{agg}}(D)$. Since the variables where the optimization takes place are different for each term of the summation, we can conclude that minimizing the summation is less than or equal to the summation of each individual minimization, thus we have $\mathcal{L}^*_{\text{agg}}(D) \le \mathcal{L}^{\text{upper}}_{\text{agg}}(D)$.

A1.2. $\mathcal{L}_{user}(D, ALPPM^*_{agg}) \le \mathcal{L}^*_{agg}(D)$

According to the property of data aggregation process, for an arbitrary but fixed $q(\boldsymbol{A}|\boldsymbol{A})$, we know that given $\boldsymbol{A}$, $\boldsymbol{A}$ is independent of $\boldsymbol{A}_m$, i.e., $\boldsymbol{A}$ is conditionally independent of $\boldsymbol{A}_m$, meaning that $\boldsymbol{A}_m$, $\boldsymbol{A}$, $\boldsymbol{A}$ form a Markov chain. Therefore, we have $I(\boldsymbol{A}_m;\boldsymbol{A}) \le I(\boldsymbol{A};\boldsymbol{A})$ for any user under the same mechanism $q(\boldsymbol{A}|\boldsymbol{A})$, which means $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg}) \le \mathcal{L}^*_{\text{agg}}(D)$.

A1.3. $\mathcal{L}^*_{user}(D) \le \mathcal{L}_{user}(D, ALPPM^*_{agg})$

Since the solution to the minimization problem in $\mathcal{L}^*_{\text{user}}(D)$ considers all possible mechanisms, which include ALPPM^*_{agg}, it must incur $\mathcal{L}^*_{\text{user}}(D)$ no larger than $\mathcal{L}_{\text{user}}(D, \text{ALPPM}^*_{agg})$.

The proof of Theorem 1 is completed.

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. ACM, pp. 1679–1688.

Andrés, M.E., Bordenabe, N.E., Chatzikokolakis, K., Palamidessi, C., 2013. Geo-indistinguishability: differential privacy for location-based systems. In: Proceedings of the 20th ACM Conference on Computer and Communications Security. ACM, pp. 901–914.

Barth, D., 2009. The Bright Side of Sitting in Traffic: Crowdsourcing Road Congestion Data. Google Official Blog

Blahut, R., 1972. Computation of channel capacity and rate-distortion functions. IEEE Trans. Inf. Theory 18 (4), 460–473.

Chan, T.-H.H., Shi, E., Song, D., 2011. Private and continual release of statistics. ACM Trans. Inf. Syst. Secur. 14 (3), 26.

Cho, E., Myers, S., Leskovec, J., 2011. Friendship and mobility: friendship and mobility: user movement in location-based social networks. In: Proc. ACM SIGKDD 2011.

Cover, T.M., Thomas, J.A., 2012. Elements of Information Theory. John Wiley & Sons.

Du, W., Zhan, Z., 2003. Using randomized response techniques for privacy-preserving data mining. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 505–510.

Dwork, C., 2008. Differential privacy: a survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, pp. 1–19.

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

Dwork, C., Naor, M., Pitassi, T., Rothblum, G.N., 2010. Differential privacy under continual observation. In: Proceedings of the Forty-Second ACM Symposium on Theory of Computing. ACM, pp. 715–724.

Erlingsson, Ú., Pihur, V., Korolova, A., 2014. Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, pp. 1054–1067.

Fan, L., Xiong, L., 2012. Real-time aggregate monitoring with differential privacy. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, pp. 2169–2173.

Ho, S.-S., Ruan, S., 2011. Differential privacy for location pattern mining. In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. ACM, pp. 17–24.

Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R., 2017. Context-aware generative adversarial privacy. Entropy 19 (12), 656.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press.

Jiang, B., Li, M., Tandon, R., 2018. Context-aware data aggregation with localized information privacy. In: 2018 IEEE Conference on Communications and Network Security (CNS). IEEE, pp. 1–9.

Kopp, C., Mock, M., May, M., 2012. Privacy-preserving distributed monitoring of visit quantities. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, pp. 438–441.

Li, Q., Cao, G., La Porta, T.F., 2014. Efficient and privacy-aware data aggregation in mobile sensing. IEEE Trans. Depend. Secure Comput. 11 (2), 115–129.

Ma, C.Y., Yau, D.K., 2015. On information-theoretic measures for quantifying privacy protection of time-series data. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. ACM, pp. 427–438.

Oya, S., Troncoso, C., Pérez-González, F., 2017. Back to the drawing board: revisiting the design of optimal location privacy-preserving mechanisms. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, pp. 1959–1972.

du Pin Calmon, F., Fawaz, N., 2012. Privacy against statistical inference. In: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, pp. 1401–1408.

Popa, R.A., Blumberg, A.J., Balakrishnan, H., Li, F.H., 2011. Privacy and accountability for location-based aggregate statistics. In: Proceedings of the 18th ACM Conference on Computer and Communications Security. ACM, pp. 653–666.

Primault, V., Boutet, A., Mokhtar, S.B., Brunie, L., 2018. The long road to computational location privacy: asurvey. IEEE Commun. Surv. Tut. 21 (3), 2772–2793.

Pyrgelis, A., De Cristofaro, E., Ross, G.J., 2016. Privacy-friendly mobility analytics using aggregate location data. In: Proceedings of the 24th ACM SIGSPATIAL international Conference on Advances in Geographic Information Systems. ACM, p. 34.

Pyrgelis, A., Troncoso, C., De Cristofaro, E., 2017a. Knock knock, who's there? Membership inference on aggregate location data. arXiv:1708.06145.

Pyrgelis, A., Troncoso, C., De Cristofaro, E., 2017. What does the crowd say about you? Evaluating aggregation-based location privacy. Proc. Priv. Enhanc. Technol. 2017 (4), 156–176.

Rastogi, V., Nath, S., 2010. Differentially private aggregation of distributed time-series with transformation and encryption. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, pp. 735–746.

Reid, R., 2011. Tomtom Admits to Sending Your Routes and Speed Information to the Police, 2011. CNET UK

Riley, P.F., 2008. The tolls of privacy: an underestimated roadblock for electronic toll collection usage. Comput. Law Secur. Rev. 24 (6), 521–528.

Sankar, L., Rajagopalan, S.R., Poor, H.V., 2013. Utility-privacy tradeoffs in databases: an information-theoretic approach. IEEE Trans. Inf. Forensics Secur. 8 (6), 838–852.

Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y., Hubaux, J.-P., 2011. Quantifying location privacy. In: 2011 IEEE Symposium on Security and Privacy. IEEE, pp. 247–262.

Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J.-P., Le Boudec, J.-Y., 2012. Protecting location privacy: optimal strategy against localization attacks. In: Proceedings of the 19th ACM conference on Computer and Communications Security. ACM, pp. 617–627.

Shokri, R., Troncoso, C., Diaz, C., Freudiger, J., Hubaux, J.-P., 2010. Unraveling an old cloak: k-anonymity for location privacy. In: Proceedings of the 9th annual ACM workshop on Privacy in the electronic society, pp. 115–118.

Sweeney, L., 2002. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (05), 557–570.
Wagner, I., Eckhoff, D., 2018. Technical privacy metrics: a systematic survey. ACM Comput. Surv. 51 (3), 1–38.
Wang, K., Chen, R., Fung, B., Yu, P., 2010. Privacy-preserving data publishing: a survey on recent developments. ACM Comput. Surv..
Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60 (309), 63–69.
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. International World Wide Web Conferences Steering Committee, pp. 1241–1250.
Zhang, W., Li, M., Tandon, R., Li, H., 2019. Online location trace privacy: An information theoretic approach. IEEE Trans. Inf. Forensics Secur. 14 (1), 235–250.

Wenjing Zhangis a Ph.D. student at the School of Cyber Engineering, Xidian University, Xi'an, China. She received both her bachelor's and master's degrees from Xidian University. She visited the Department of Electrical and Computer Engineering at the University of Arizona from 2016 to 2018. Her current research interests focus on data privacy, privacy metrics and machine learning.

Bo Jiang is a research assistant and a full-time Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Arizona. He received his master's degree from Worcester Polytechnic Institute and his bachelor's degree and master's degree from Harbin Institute of Technology. His current research interests include security and privacy-protection, machine learning and image processing.

Ming Li is an Associate Professor in the Department of Electrical and Computer Engineering of University of Arizona. He was an Assistant Professor in the Computer Science Department at Utah State University from 2011 to 2015. He received his Ph.D. in ECE from Worcester Polytechnic Institute in 2011. His main research interests are wireless networks and security, with current emphases on wireless network optimization, wireless security and privacy, and cyber-physical system security. He received the NSF Early Faculty Development (CAREER) Award in 2014, and the ONR Young Investigator Program (YIP) Award in 2016. He is a senior member of IEEE and a member of ACM. NSF Early Faculty Development (CAREER) Award in 2014, and the ONR Young Investigator Program (YIP) Award in 2016. He is a senior member of IEEE and a member of ACM.

Ravi Tandon is an Assistant Professor in the Department of ECE at the University of Arizona. Prior to joining the University of Arizona in Fall 2015, he was a Research Assistant Professor at Vir- ginia Tech with positions in the Bradley Department of ECE, Hume Center for National Security and Technology and at the Discovery Analytics Center in the Department of Computer Science. He received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Kanpur (IIT Kanpur) in 2004 and the Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park (UMCP) in 2010. From 2010 to 2012, he was a post-doctoral research associate at Princeton University. He is a recipient of the 2018 Keysight Early Career Professor Award, NSF CAREER Award in 2017, and a Best Paper Award at IEEE GLOBECOM 2011. He is a Senior Member of IEEE and currently serves as an Editor for IEEE Transactions on Wireless Communications and IEEE Transactions on Communications. His current research interests include information theory and its applications to wireless networks, communications, security and privacy, machine learning and data mining.

Qiao Liu is a Lecturer in the School of Cyber Engineering at the Xidian University. He his received Ph.D. degree from Xidian University in 2017, and received his B.S. degree from Xidian University in 2011. From 2014 to 2016, he was a visiting Ph.D. student at the department of ECE at the University of Waterloo, Canada, supported by CSC. His research interests include wireless network security, physical layer security, and cooperative communication.

Hui Li is a Professor at the School of Cyber Engineering, Xidian University, Xi'an, China. He received B.Sc. degree from Fudan University in 1990, M.Sc. and Ph.D. degrees from Xidian University in 1993 and 1998. In 2009, he was with Department of ECE, University of Waterloo as a visiting scholar. His research interests are in the areas of cryptography, privacy computing, wireless network security, information theory.