Future Generation Computer Systems 86 (2018) 1513–1522

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

On the limitations of existing notions of location privacy

Kai Dong [a,*], Taolin Guo [a], Haibo Ye [b], Xuansong Li [c], Zhen Ling [a]

[a] School of Computer Science and Engineering, Southeast University, China
[b] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China
[c] School of Computer Science and Engineering, Nanjing University of Science and Technology, China

ARTICLE INFO

Article history:
Received 14 January 2017
Received in revised form 21 April 2017
Accepted 20 May 2017
Available online 15 July 2017

Keywords:
Location privacy
Differential privacy
Obfuscation

ABSTRACT

In the context of a single report of location information, existing researches define location privacy by adversary's uncertainty, inaccuracy, or incorrectness of the estimation, or by geo-indistinguishability which is a generalization of differential privacy. Each of these existing notions has problems in some specific scenarios. In this paper we illustrate the limitations of existing notions by constructing such scenarios, and introduce a formal definition on location privacy by quantifying the distance between the prior and posterior distribution over the possible locations. Further more, we show how to construct a near-optimal obfuscation mechanism by solving an optimization problem. We compare our proposed mechanism with the Laplace noise based geo-indistinguishable mechanism, and Shokri's optimal obfuscation mechanism, using both our proposed privacy metric and the traditional metric based on the estimated distance errors. The results show that our proposed metric better describes location privacy and our proposed mechanism makes a better tradeoff between privacy and utility.

1. Introduction

With the development of mobile computing and the wide spread of mobile devices, mobile social network applications have become increasingly prevalent across mobile users. These applications rely on various location based services (LBS) to make use of users' location information and thus can provide users personalized services. Without an adequate location privacy preserving mechanism, users may be hesitant to use these applications.

Researchers have proposed a variety of location privacy preserving mechanisms (LPPM) which allow users to make use of the LBS with reduced location information [1–4]. These LPPMs provide different trade-offs between location privacy and LBS utility, offering alternatives to better meet individual requirements of different users.

However the comparison between LPPMs can be tricky due to a lack of reasonable privacy benchmark for location information. For given datasets and adversary assumptions, many early researches define user's privacy [2] by the "'uncertainty" of the adversary, which tells the probability that the adversary will make a wrong estimate. Then "incorrectness", which is a combination of uncertainty and "inaccuracy", is introduced as a better definition of privacy [5], since privacy deals with not only the probability of an error, but also the magnitude of this error. This notion of privacy is reasonable, however it is difficult to measure the error magnitude, since no distance metric can be used for all situations. For example, one may suggest using the Euclidean distance, and then he will find that with a same distance, the privacy that the real location and the estimated location belong to a same region, can be different with that they belong to two different regions.

In recent years, differential privacy [6] gains popularity since it abstracts from the side information of the adversary. In the context of location privacy, geo-indistinguishability is introduced and supposed to be "independent of the prior". It defines privacy by the maximum difference among the probability of reporting a location from all possible real locations (and this difference decays with the distance between two possible real locations). This definition can be problematic if the prior is taken into account, and we prove in this paper that geo-indistinguishability is not independent of the prior, instead it is based on the assumption that the prior is unknown.

In this paper, we propose DPLO (short for differentially private location obfuscation) as a notion of location privacy by describing the difference between the prior and the posterior knowledge of the adversary. We distinguish the prior distribution over locations before and that after the user decides to access LBS at some location, and use the latter as the prior knowledge. This is because the user decides to trade-off privacy for utility, and no matter which LPPM is used, the decline of privacy is inevitable. Moreover, we distinguish the posterior distribution over locations before and that after the LPPM finally outputs some obfuscated location, and use the former as the posterior knowledge. This is because

* Corresponding author.
E-mail addresses: dk@seu.edu.cn (K. Dong), guotaolin@seu.edu.cn (T. Guo), yhb@nuaa.edu.cn (H. Ye), lixs@njust.edu.cn (X. Li), zhenling@seu.edu.cn (Z. Ling).

http://dx.doi.org/10.1016/j.future.2017.05.045
0167-739X/

our goal is not to quantify privacy for some specific outputs of an LPPM, but to propose a general privacy metric for evaluating LPPMs based on a probabilistic model. We study the problem of optimizing the DPLO under given quality constraints, and construct a near-optimal obfuscation mechanism by solving a non-linear optimization problem. Our proposed obfuscation mechanism is evaluated by comparison with selected existing mechanisms based on the same datasets used by other literatures.

2. Existing notions of location privacy

In situations when people do not have to disclose their locations, they can use security approaches such as encryption or some other access control mechanism to ensure location privacy. However, in most other situations, one has to trade-off between location privacy and utility, e.g., when a user accesses an untrustworthy LBS. We focus on location privacy when such trade-offs take place, which is so-called the computational location privacy introduced by Krumm [7].

Since location privacy is computational, we can make comparison between different privacy preserving mechanisms. During the last decade, a variety of privacy metrics have been proposed and they mainly fall into three categories [4]: *k*-anonymity, expected distance error, and differential privacy.

2.1. Uncertainty: k-anonymity and location entropy

k-anonymity [8] is a property of anonymized data in databases. Briefly speaking, a table satisfying *k*-anonymity means that, for each record in this table, there exist at least $k-1$ other records with exactly the same quasi-identifier (sensitive attribute) values. To achieve *k*-anonymity, generalization techniques are often used. This notion is widely adopted in early researches on location privacy. Gruteser et al. [1] introduce a *cloaking* based mechanism, which employs a trusted anonymizer (or uses a peer-to-peer algorithm to solve the one point of failure problem as in [9]), to aggregate location reports from at least *k* users (or *dummy* users to achieve a higher *k* in sparse regions as in [10]), and replace the locations with one generalized area to ensure *k*-anonymity. Beresford et al. [2] introduce a *confusion* based mechanism. The idea behind confusion is that if at least *k* users change their pseudonyms and report a same generalized location at the same time (e.g., two cars at a crossroad in path confusion [10]), they become indistinguishable since then. Other techniques like *cache* [11] may also be used to better trade-off between location privacy and utility, while the notion of privacy remains the same: location privacy is defined as the **uncertainty** of adversary, and the more possible locations there are, the higher location privacy will be.

2.2. Incorrectness: expected distance error

Suppose an LPPM which outputs obfuscated locations based on user's real locations. It is obvious that if the distance between the real location and the obfuscated location is not large enough, these two locations can belong to a same logical location, e.g., two different locations in a hospital. In this case, the user's location privacy is not preserved. The minimal distance which ensures two locations each belongs to a different logical location can vary widely accordingly. It depends highly on the map information, the scale of location, the type of application, the privacy requirement of the user, and many other contextual information, and is too complex to define. An intuitive way to improve privacy is to make larger the distance between the real location and the obfuscated location, since the larger this distance is, the less likely the two locations will belong to a same logical location. In the meanwhile, this distance contributes to the quality loss, so the complex problem of trading-off between privacy and utility can be transformed to the problem of deciding the distance error.

However, this notion of location privacy is still problematic. In particular, a smart adversary may compute an estimated location based on his knowledge of the obfuscation algorithm used by the LPPM. So the expected distance error should measure the distance between the real location and the estimated location, instead of the reported one. Further more, a smart LPPM should consider the adversary's knowledge and capability to better trade-off between privacy and utility. Shokri et al. [3,5] introduce a comprehensive location privacy notion by completing the existing adversary's model based on the understanding that the privacy of users and the success of the adversary are two sides of the same coin. Unlike traditional *k*-anonymity based approaches, [5] measures location privacy using so-called **incorrectness**, which is a combination of the adversary's *uncertainty* and *inaccuracy* on the estimated locations.

2.3. Differential privacy: geo-indistinguishability

Differential privacy [6] is a notion of privacy from the area of statistical databases. It is introduced to protect against deanonymization techniques which identify personal information by linking two or more separately anonymized databases. It can be used to measure location privacy in statistical databases [12,13]. However, location privacy in LBS scenarios is to some extent different, since most LBSs require specific location information of a single user instead of some statistics on aggregate information of multiple users. Dewri [14] proposes differential perturbation which is a hybrid of differential privacy and *k*-anonymity. In this approach, the *k* locations in an anonymity set are required to have similar probabilities to report a same obfuscated location.

Geo-indistinguishability [4,15] proposed by Andrés and Bordenabe et al., has gained popularity in recent years. It relaxes Dewri's constraint of putting locations in anonymity sets. Multivariate Laplace noise is used by Andres et al. [4] to achieve ϵ-geo-indistinguishability. Then in their later work, Bordenabe et al. [15] propose an optimal geo-indistinguishable mechanism by solving a linear optimization problem, which chooses obfuscation probability distribution function $f(\cdot,\cdot)$ (i.e., to choose the noise distribution instead of simply using Laplace noise), to minimize the service quality loss.

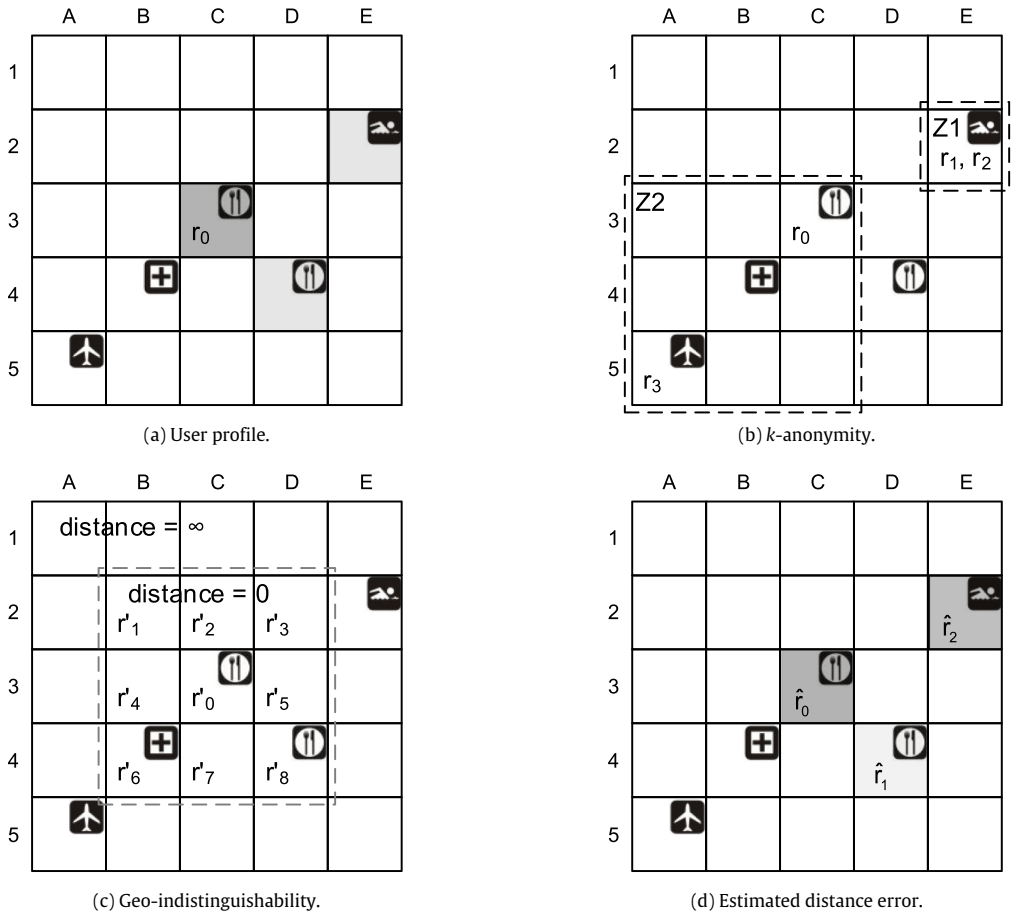
Geo-indistinguishability is now widely adopted [16,17], and it is also improved in recent approaches by considering the temporal correlations of multiple locations [18,19].

3. Limitations of existing notions

The limitations of existing notions motivate this work. For better understanding these limitations, we illustrate them in detail by computing location privacy using the formal definitions of existing notions under the following scenario settings and adversary assumptions. In Table 1, we summarize the main notations introduced throughout this article.

3.1. Scenario description and adversary assumptions

The example scenario is as shown in Fig. 1. We focus on a 5×5 grid consisting of 25 square regions with each of which represents a location. The symbol in the bottom left corner of a region indicates whether this location is a real location r, an obfuscated locations r' or an estimated locations $\hat{r}$. We use $\mathcal{R}$ to represent the set of all possible real locations (it is obvious that $\hat{r} \in \mathcal{R}$, if $\mathcal{R}$ is known by the adversary), and $\mathcal{R}'$ to represent the set of all possible obfuscated locations. The icon in the top right corner



(a) User profile.

(b) k-anonymity.

(c) Geo-indistinguishability.

(d) Estimated distance error.

Fig. 1. Limitations of existing notions: (a) User profile, probability distribution on user's location. (b) Two anonymity zones with the same k and the same location entropy but not equally private. (c) With $\epsilon = 0$, there is still a 4/9 chance of reporting r'_1, r'_2, r'_4 or r'_6, each of which identifies user's real location r. (d) When the estimated distance error increases, the probability of identifying the real location can also increase (i.e., privacy can degrade).

Table 1
Summary of notations.

Symbol	Meaning
$r, r', \hat{r}$	Real location, obfuscated/reported location, estimated location
$\mathcal{R}, \mathcal{R}'$	The set of all real locations, the set of all obfuscated locations
$\psi(r)$	User profile (probability of being at location r)
$f(r'\|r)$	Obfuscation function (probability of reporting r')
$h(\hat{r}\|r')$	Attack function (probability of estimating $\hat{r}$ as real location)
$Q(r'\|r)$	Utility of reporting r' instead of r
$Q_{\min}$	User's minimum acceptable utility
$\mathcal{A}(\hat{r}\|r)$	Prior probability of estimating location $\hat{r}$
$\mathcal{B}(\hat{r}\|r)$	Posterior probability of estimating location $\hat{r}$
$d_{\mathcal{A}}(r', r)$	Distortion function of two locations (considering utility)

of a region indicates its logical location, e.g., a restaurant, a gym or a hospital, etc.

Suppose our target user u_0 is now locating at $r_0 = C3$, and his profile can be modeled by a prior distribution ψ on $\mathcal{R}$ is as shown in Fig. 1(a), with the grayness of a region shows the prior probability that the user is located at this location. In our scenario, we have $\psi(C3) = 0.5$, $\psi(D4) = \psi(E2) = 0.25$.

We assume the adversary is aware of the LPPM's internal algorithm and the user profile (the adversary can accumulate this knowledge with repeated observations/eavesdropping). Base on these assumptions, we show the gap between the existing notions of location privacy, and what privacy naturally is.

3.2. Limitations of k-anonymity

By applying Shannon's classic measure of entropy, Beresford et al. [2] define location privacy as location entropy:

$$\text{Privacy} = -\sum_{r} \Pr(r)\log_2 \Pr(r) \tag{1}$$

where $\Pr(r)$ represents the probability that location r is the real location in the adversary's estimate. Ideally, $k = 2^b$ locations with exactly the same probability result in a location entropy that equals to b.

The limitations to this notion includes:

1. k-anonymity suffers from homogeneity attacks and background knowledge attacks [20].
2. Uncertainty of the adversary does not always mean privacy of the users [5].

These limitations of k-anonymity have been widely discussed, here we give a simple example for illustration. As shown in Fig. 1(b), four users $u_0, \ldots, u_3$ locate at $r_0, \ldots, r_3$ respectively. To achieve 2-anonymity, r_1 and r_2 are aggregated and generalized to zone $Z1$, while r_0 and r_3 are aggregated and generalized to zone $Z2$. Although users in both zones have the same location entropy ($k = 2$), location privacy in zone $Z2$ is obviously preserved much better than in $Z1$. In our example, since locations of two anonymous users are close to each other (r_1 and r_2), adversary can deduce that both users locate at a logical location (the gym at $D2$) by performing a homogeneity attack. Moreover, if the adversary

is aware of u_0's profile, he can deduce that u_0 locates at $C3$ by performing a background knowledge attack.

3.3. Limitations of geo-indistinguishability

Geo-indistinguishability is a generalization of differential privacy in the context of location privacy. In [4], ϵ-geo-indistinguishability is defined as

$$d_{\mathcal{P}}(f(r'|r), f(r'|\hat{r})) \leq \epsilon d_{\mathcal{R}}(r, \hat{r}) \tag{2}$$

where r, r' and $\hat{r}$ are arbitrary locations, $f(r', \cdot)$ is a probabilistic function for selecting the obfuscated location r', and $d_{\mathcal{P}}(\cdot, \cdot)$ measures the supremum of distance between two distributions, $d_{\mathcal{R}}(\cdot, \cdot)$ measures distance between two locations.

The limitations to this notion of location privacy include:

1. ϵ-geo-indistinguishability as a notion of location privacy may be problematic, since privacy is not always enhanced with ϵ decreases.
2. ϵ-geo-indistinguishability does not perform well for location traces, since privacy degrades rapidly when traces become longer. Some recent approaches introduce notions of location privacy considering temporal correlations [18,19]. However we are only interested in location privacy at a single time-stamp in this paper, and we leave the extension of location privacy for traces as future work.

In Formula (2), $d_{\mathcal{R}}(\cdot, \cdot)$ is a metric measuring the distance between two locations. In the vanilla geo-indistinguishability [4], the euclidean distance is used; and in optimal geo-indistinguishability [15], this metric is defined accordingly in various applications. Suppose the following distance metric,

$$d_{\mathcal{R}}(r_1, r_2) = \begin{cases} 0 & \text{if } r_1 \in \text{Grid}_{3\times3}(r_2) \\ \infty & \text{otherwise} \end{cases} \tag{3}$$

where $\text{Grid}_{3\times3}(r_2)$ is a r_2-centered, 3×3 grid consisting of 9 regions. As in Fig. 1(c), the dotted-line square can be represented as $\text{Grid}_{3\times3}(r)$. This metric is reasonable for many LBSs, e.g., which provides a user with nearby geographic information based on his location r. If the reported location r' is within an area, i.e., $r' \in \text{Grid}_{3\times3}(r)$, the utility of the LBS can be ensured, so the distance is set to 0. Otherwise, the information obtained by the user may be completely useless, so the distance is set to ∞.

With this metric, we can construct a simple mechanism satisfying 0-geo-indistinguishability as follows:

$$f(r'|r) = 1/9, \quad \forall r \in R, r' \in \text{Grid}_{3\times3}(r). \tag{4}$$

Note that in statistical databases, 0-differential privacy means "complete privacy" since it ensures that the personal information will never be identified. However a 0-geo-indistinguishability mechanism may perform poorly in privacy. For example the 0-geo-indistinguishability mechanism in Eq. (4), the user will have a 4/9 chance to report an obfuscated location r'_1, r'_2, r'_4 or r'_6, and reporting any of these locations leads to disclosure of user's real location.

Even if we use some other distance metric such as the Euclidean distance, and use some complicated noise distribution such as Laplace noise as in [4] to construct some other geo-indistinguishability mechanism, there is still great difference between location privacy and geo-indistinguishability.

One may argue that geo-indistinguishability, just like differential privacy, is designed for situations when priors are unknown. This statement is true for the differential privacy in statistical databases, since it guarantees that the posterior probability is as much as the prior probability, so we say it is "independent of the prior". However, there is a misunderstanding for the geo-indistinguishability, since it is no independent of the prior, instead it makes a strong hypothesis that the prior cannot be known.

3.4. Limitations of optimal geo-indistinguishability

Now let us move on to the optimal geo-indistinguishability proposed in [15]. It employs the definition of vanilla geo-indistinguishability [4] as the notion of location privacy, and is essentially a better trade-off between utility and geo-indistinguishability.

Since geo-indistinguishability is problematic, we can also conclude that the optimal geo-indistinguishability is also problematic. Back to our example shown in Fig. 1(c). We can prove that the only mechanism achieves $\epsilon = 0$ is the mechanism we proposed in Eq. (4), so this mechanism is also an optimal 0-geo-indistinguishable mechanism, and we have already proven this mechanism performs poorly in privacy.

3.5. Limitations of expected distance error

Using the estimated distance error, location privacy is defined as:

$$\text{Privacy}(\psi, f, h, d_P) = \sum_{r,r',\hat{r}} \psi(r) f(r'|r) h(\hat{r}|r') d_P(\hat{r}, r) \tag{5}$$

where $\psi(r)$ represents the prior probability (i.e., the user profile) that the user locates at location r, $f(r'|r)$ represents the probability that the LPPM outputs an obfuscated location r' based on an real location r, $h(\hat{r}|r)$ represents the probability that the adversary guesses the user's location to be $\hat{r}$ based on a reported location r', and $d_P(\hat{r}, r)$ represents the distance between locations $\hat{r}$ and r.

In [3], Shokri introduces an optimal strategy for location privacy by solving a linear program, which chooses LPPM's obfuscation probability distribution function $f(\cdot, \cdot)$, to maximize the location privacy defined in Eq. (5), subject to service quality constraints which give an upper bound on the distance between real location r and obfuscated location r'. This approach is optimal with this notion, however, there still remain two limitations:

1. It relies on the modeling of adversary's side information, meaning that it suffers from background knowledge attacks.
2. The real location privacy is not always enhanced with estimated distance error increases.

With this notion, if no location is reported, the expected distance error can be computed based on the user profile as:

$$\text{Privacy}_{\text{Prior}} = \sum_{\hat{r} \in R} \psi(\hat{r}) d(r, \hat{r}) \approx 0.91,$$

where $d(\cdot, \cdot)$ here is the euclidean distance.

Assume a mechanism outputs some obfuscated location to report (e.g., when $f(D2|C3)$ is relatively high), and finally makes the adversary estimates a posterior distribution as shown in Fig. 1(d): $\Pr(C3) = 0.5$, $\Pr(E2) = 0.45$ and $\Pr(D4) = 0.05$. The expected distance error can be computed as:

$$\text{Privacy}_{\text{Posterior}} = \sum_{\hat{r} \in R} \Pr(\hat{r}) d(r, \hat{r}) \approx 1.08.$$

It is weird that the privacy increases with this location report. Shokri [3] uses Eq. (5) to compute the averaged minimum posterior privacy on all locations $r \in R$, so the privacy with posterior is always lower than the privacy with only prior. However, for a single user's single access to an LBS, the case as in our example can happen occasionally.

This problem becomes especially pronounced if we take into account other side information. Suppose the user reports his location at lunch time, the adversary can deduce that this user is unlikely to be at location $E2$, since $E2$ is a gym and nobody will do strenuous exercises after a meal. With only user profile, the adversary has a

2/3 chance to identify user's real location; while with posterior as shown in Fig. 1(d), the chance increases to 10/11. This example shows that privacy is not always enhanced with the estimated distance error increases, and also the importance of independence of the prior.

4. DPLO: differentially private location obfuscation

We have shown the limitations of using uncertainty, inaccuracy, incorrectness, and geo-indistinguishability as the notion of location privacy. Our goal is to provide a formal notion of location privacy according to the common understanding that privacy is "the ability of an individual to seclude information about himself". In access to an LBS, the user has to trade-off between location privacy and utility, and report some obfuscated location information. In this context, complete privacy means "adversary knows no additional information from the obfuscated location". Based on this understanding, we introduce differentially private location obfuscation (DPLO) as a formal notion of location privacy, by quantifying the additional knowledge probably disclosed by the obfuscation. Some important notations are listed in Table 1.

4.1. Assumptions

Let $\mathcal{R}$ be a set of points of interest, including all possible real locations of a given user, and $\mathcal{R}'$ be a set of locations, including all possible reported locations. Suppose the user locates at location $r \in \mathcal{R}$, he uses an obfuscation mechanism to protect his location. This obfuscation mechanism chooses a pseudo-location $r' \in \mathcal{R}'$ by sampling from a probability distribution $f(r'|r)$.

The adversary knows the obfuscation function $f(\cdot,\cdot)$, and he also knows in prior the user profile $\psi(r_i)$, i.e., the probability distribution of user's real location. Based on any obfuscated location r' he obtains, he will estimate a location $\hat{r} \in \mathcal{R}$ as the user's real location. The attack function he uses can be represented as $h(\hat{r}|r')$. Typically, a Bayesian adversary uses Bayesian inference attack on the obfuscation mechanism, thus he can estimate $\hat{r}$ for each observed r', with prior information ψ:

$$h(\hat{r}|r') = \frac{\Pr(\hat{r}, r')}{\Pr(r')} = \frac{f(r'|\hat{r})\psi(\hat{r})}{\sum_r f(r'|r)\psi(r)}. \tag{6}$$

Note that, due to utility reasons, r' is always around r, so $h(\hat{r}|r')$ is always no less than $\psi(\hat{r})$.

4.2. Quality metric

The quality of the LBS for a user locating at r reports r' can be computed as:

$$Q(r', r) = e^{-d_{\mathcal{R}}(r'|r)} \tag{7}$$

where the distance function $d_{\mathcal{R}}(\cdot,\cdot)$ can be defined as different metrics accordingly. For example, the Euclidean distance $d_E(\cdot,\cdot)$ between the locations is a typical metric.

For any given real location r, the LPPM should ensure that the utility of reporting an obfuscated location r', so we have:

$$Q(r', r) \geq Q_{\min}. \tag{8}$$

4.3. Definition of posterior distribution

For a given real location r, the probability that the adversary will guess that the location to be $\hat{r}$ is:

$$\mathcal{B}(\hat{r}|r) = \sum_{r'} f(r'|r)h(\hat{r}|r'). \tag{9}$$

Typically for a Bayesian adversary, we have

$$\mathcal{B}(\hat{r}|r) = \sum_{r'} \frac{f(r'|r)f(r'|\hat{r})\psi(\hat{r})}{\sum_{r_i} f(r'|r_i)\psi(r_i)}. \tag{10}$$

4.4. Definition of prior distribution

The definition on "prior distribution" worth taking up analysis. We start with the clarification of two different moments. The first moment is the time before the user decides to use the LBS. The probability distribution prior to this moment is no doubt $\psi(r)$. The second moment is the time after the user has decided to use the LBS at some real location r, but before he really report some location r'. At this moment, the user has decided to trade-off privacy for utility, and the privacy will definitely degrade if the utility is considered, no matter what obfuscation mechanism is used. So the probability distribution prior to this moment is no longer the user profile $\psi(r)$, and we use $\mathcal{A}(\hat{r}|r)$ to denote this prior, which can be computed as follows:

$$\mathcal{A}(\hat{r}|r) = \frac{d_A(\hat{r}, r)\psi(\hat{r})}{\sum_{r_i} d_A(r_i, r)\psi(r_i)}, \tag{11}$$

where d_A is another metric different from $d_{\mathcal{R}}$. If the distance metric $d_{\mathcal{R}}$ is defined by the Euclidean distance d_E, we can compute d_A as:

$$d_A(\hat{r}|r) = e^{-d_{\mathcal{R}}(\hat{r}|r)/2}. \tag{12}$$

No matter which obfuscation mechanism is chosen, we have that $\mathcal{A}(\hat{r}|r)$ is an upper bound (but not necessarily a supremum) of location privacy in case that utility is ensured. For better understanding the difference between the prior ψ and $\mathcal{A}$, we give an example: Suppose a user's real location is somewhere in Los Angeles, and his profile shows that there is a 50% chance that he is in Los Angeles, and another 50% chance in San Francisco. The overall prior ψ is a distribution mapping from all locations in both cities, and we have $\sum_{r\in LA}\psi(r) = \sum_{r\in SFO}\psi(r) = 0.5$. Now this user accesses an LBS, which requires city-level accuracy of location, so he can only report some pseudo-location in the same city. In this case, the function $\mathcal{A}$ indicates the prior distribution ensuring utility which maps from the locations in only one city, and we have $\sum_{r\in LA}\mathcal{A}(r) = 1$ and $\sum_{r\in SFO}\mathcal{A}(r) = 0$. This is because the user has to trade-off privacy for utility, and at the time he decides to report a location for some utility, his location privacy degrades.

4.5. Definition of DPLO

We define ϵ-DPLO, which is short for "differentially private location obfuscation", as the notion of privacy by quantifying the difference between the prior and the posterior of the adversary.

Definition 1 (*DPLO*)**.** Let ϵ be a positive real number, a location obfuscation mechanism $f(\cdot|\cdot)$ satisfies ϵ-DPLO iff for all r, $\hat{r}$:

$$e^{-\epsilon} \leq \frac{\mathcal{A}(\hat{r}|r)}{\mathcal{B}(\hat{r}|r)} \leq e^{\epsilon} \tag{13}$$

where $\mathcal{A}(\hat{r}|r)$ represents the prior distribution which is a constant for any given pair of r and $\hat{r}$ and is defined in Eq. (11), and $\mathcal{B}(\hat{r}|r)$ represents the posterior distribution which is a function of $f(\cdot|\cdot)$ defined in Eq. (9).

4.6. Examples on computing ϵ-DPLO

In Fig. 1 we provide several sample obfuscation mechanisms to illustrate the limitations of existing notions of location privacy. Here, we compute the ϵ-DPLO that each of these mechanisms satisfies.

For all the cases, we assume a Bayesian adversary to compute the posterior distribution by Eq. (10), and suppose the user profile $\psi(r)$ is as shown in Fig. 1(a): $\psi(C3) = 0.5$, $\psi(D4) = \psi(E2) = 0.25$. The distortion metric $d_A(\cdot,\cdot)$ can be computed from the distance metric $d_{\mathcal{R}(\cdot,\cdot)}$ we assumed in Eq. (3), so we have:

$$d_A(r_1, r_2) = \begin{cases} 1 & \text{if } r_1 \in \text{Grid}_{5\times5}(r_2) \\ 0 & \text{otherwise} \end{cases}$$

where $\text{Grid}_{5\times5}(r_2)$ is a r_2-centered, 5 $\times$ 5 grid. The prior distribution $\mathcal{A}(\hat{r}|r)$ can be computed by Eq. (11), so we have $\mathcal{A}(\hat{r}|C3) = \mathcal{A}(\hat{r}|D4) = \mathcal{A}(\hat{r}|E2) = \psi(\hat{r})$, for all $\hat{r} \in R$.

Now we can compute ϵ for each sample mechanism. For the k-Anonymity Mechanism we assumed in Section 3.2, it satisfies ∞-DPLO; for the 0-Geo-Indistinguishable Mechanism we assumed in Section 3.3 (Eq. (4)), it satisfies 1.35-DPLO; for the estimated distance error based mechanism we assumed in Section 3.5, it satisfies 1.61-DPLO.

Suppose an obfuscation mechanism $f_{\text{opt}}(D3|r) = 1$, for all $r \in R$. We have $f_{\text{opt}}(\cdot|\cdot)$ satisfies 0-DPLO. This is reasonable, since the adversary will obtain no extra knowledge from this obfuscation, and we say this mechanism is optimal, since it ensures utility and privacy simultaneously.

5. Near-optimal DPLO mechanism

In the previous section, we introduce DPLO as the notion of location privacy, and show via case study how to compute the ϵ-DPLO that various obfuscation mechanism can satisfy. In this section, we propose a method of constructing a near-optimal mechanism by solving an optimization problem.

5.1. Problem statement

Given the distortion function $d_{\mathcal{R}}(\cdot,\cdot)$, and the user profile $\psi(\cdot)$ on a set of locations $\mathcal{R}$ as prior knowledge, the problem is finding the obfuscation function $f(\cdot,\cdot)$ that minimizes the chance of identifying the real locations, i.e., minimizes ϵ as in Definition 1. The solution must consider that the adversary is aware of the obfuscated location r' and the obfuscation function $f(\cdot,\cdot)$.

5.2. Optimal mechanism

With inequality constraint (15) in Definition 1, we can construct an optimal obfuscation mechanism for a given set $\mathcal{R}$ of all possible real locations and a given set $\mathcal{R}'$ of all possible reported locations, by solving a nonlinear optimization problem:

Choose $f(r'|r)$ in order to

$$\textbf{Min} \quad \epsilon \tag{14}$$

$$\textbf{s.t.} \quad e^{-\epsilon} \le \frac{\mathcal{A}(r,\hat{r})}{\mathcal{B}(r,\hat{r})} \le e^{\epsilon}, \quad \forall r, \hat{r} \in \mathcal{R} \tag{15}$$

$$\Sigma_{r'} f(r'|r) = 1, \quad \forall r \in \mathcal{R} \tag{16}$$

$$f(r'|r) \ge 0, \quad \forall r' \in \mathcal{R}', \ r \in \mathcal{R} \tag{17}$$

$$\epsilon \ge 0, \tag{18}$$

$$e^{-d_{\mathcal{R}}(r'|r)} \ge Q_{\min}, \quad \text{if } f(r'|r) > 0, \forall r' \in \mathcal{R}', r \in \mathcal{R}. \tag{19}$$

The inequality constraint (15) and (18) can be combined and transformed to:

$$(\ln \mathcal{A}(r,\hat{r}) - \ln \mathcal{B}(r,\hat{r}))^2 \le \epsilon^2.$$

Algorithm 1: Find Near-Optimal Mechanism

```
Input: loop
Output: f
y_gmin ← ∞;
x, x' ← zeros(|R'|, |R|);
for a = 0; a < loop; a++ do
    for b in Range(x) do
        x_b ← Random();
    end
    Compute y(x) by function 21;
    Compute arg min_x y(x), y_lmin by sub-gradient method;
    if y_lmin < y_gmin then
        y_gmin ← y_lmin;
        x' ← x;
    end
end
Compute f(x') by Equation 20;
return f;
```

Minimizing ϵ while $\epsilon \ge 0$ is equivalent to minimizing

$$\max_{r,\hat{r}} (\ln \mathcal{A}(r,\hat{r}) - \ln \mathcal{B}(r,\hat{r}))^2.$$

By doing so, the variable ϵ is reduced.

For any given r and $\hat{r}$, $\mathcal{A}(r,\hat{r})$ is a constant, and $\mathcal{B}(r,\hat{r})$ is a function of variables $f(\cdot|\cdot)$. These variables are non-independent, e.g., $f(r'_1|r)$ and $f(r'_2|r)$ subject to equality constraint (16). We assume a set of independent non-negative integers $x_{i,j}$ with $i \in \mathcal{R}', j \in \mathcal{R}$, and let

$$f_x(r', r) \triangleq x_{r',r} / \sum_i x_{i,r}. \tag{20}$$

So we have $f_x(r', r) = f(r'|r)$. Using $f_x(\cdot,\cdot)$ to replace $f(\cdot|\cdot)$ in $\mathcal{B}(\cdot|\cdot)$, the optimal mechanism can be constructed by solving the following optimization problem:

Choose $x_{i,j}$ in order to

$$\textbf{Min} \quad y(x) = \max_{r,\hat{r}} \left(\ln \mathcal{A}(r,\hat{r}) - \ln \sum_{r' \in \mathcal{R}'} \frac{f_x(r',r) f_x(r',\hat{r}) \psi(\hat{r})}{\sum_{k \in \mathcal{R}} f_x(r',k) \psi(k)} \right)^2 \tag{21}$$

$$\textbf{s.t.} \quad (e^{-d_{\mathcal{R}}(r'|r)} - Q_{\min}) \cdot x_{r',r} \ge 0, \quad \forall r' \in \mathcal{R}', r \in \mathcal{R} \tag{22}$$

$$x_{r',r} \ge 0 \quad \forall r' \in \mathcal{R}', r \in \mathcal{R}. \tag{23}$$

5.3. Near-optimal mechanism

This optimization may have many local minimums. To find the approximate global minimum, a typical way is to find several local minimums with random initialization of independent variables by sub-gradient method as shown in Algorithm 2, and treat the smallest local minimum as the approximate global minimum as shown in Algorithm 1. The algorithm takes *loop* different random initialization, and finds x that leads to the smallest local minimum, and finally construct the near-optimal mechanism. The larger *loop* is, the more likely that the near-optimal mechanism is optimal.

6. Experiment and evaluation

In this section, we evaluate our near-optimal mechanism and compare it to some existing LPPMs.

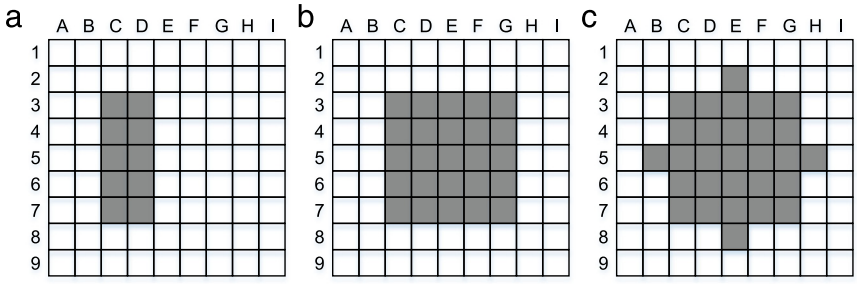


Fig. 2. Priors considered.

Algorithm 2: Sub-Gradient Method

Input: $\Delta = 0.0001, x, y(x)$
Output: x^{opt}, y_{lmin}
$flag \leftarrow true$;
$x^{opt} \leftarrow x$;
while *flag* **do**
  $flag \leftarrow false$;
  for *i in Range*(x) **do**
    $x^L, x^R \leftarrow x^{opt}$;
    $x_i^L \leftarrow x_i^L - \Delta$;
    $x_i^R \leftarrow x_i^R + \Delta$;
    if $y(x^L) < y(x^{opt})$ **then**
      $x_i^{opt} \leftarrow x_i^{opt} - \Delta$;
      $flag \leftarrow true$;
    end
    else
      if $y(x^R) < y(x^{opt})$ **then**
        $x_i^{opt} \leftarrow x_i^{opt} + \Delta$;
        $flag \leftarrow true$;
      end
    end
  end
end
return $x^{opt}, y(x^{opt})$;

6.1. Experimental method

The selected compared LPPMs and benchmarks are as follows.

Compared LPPMs. We choose two mechanisms for comparison.

- Shokri's optimal obfuscation mechanism presented in [3]. Here optimal means it achieves the maximum estimated distance error.
- Andrés' geo-indistinguishable mechanism presented in [4]. This mechanism adds the Laplace noise to the location. We do not use the optimal geo-indistinguishable mechanism presented in [15], since it assumes the adversary is aware of the user profile, and under this assumption, some other mechanism performs better in privacy, e.g., the previous mechanism.

Comparison with simple cloaking/obfuscation mechanisms is meaningless, since they are proven to perform poorly according to [3,4].

Datasets. Two datasets are used.

- The simulated data in [4], which considers three different profiles. Simulation on this data is simple and also straightforward.
- The GeoLife GPS Trajectories dataset [21], which contains 17,621 trajectories of 182 users from April 2007 to August 2012. The majority of the data was created in Beijing, China.

Privacy metrics. Two privacy metrics are used.

- The estimated distance error as defined in Eq. (5), this metric is used by most of the recent location privacy literatures.
- DPLO by Definition 1, our proposed metric.

6.2. Experiments on simulated profiles

The simulated user profiles are as shown in Fig. 2, in each case, the probability distribution is accumulated in the regions in the gray area, and distributed uniformly over them.

We use the same settings of Andrés' mechanism as in [4], and the quality loss is 107.30 m.[1] This quality loss value comes from a simple cloaking mechanism with a fixed quality loss which always reports the center of 3×3 anonymity zone. We fix this value for all our selected compared mechanisms, and especially we let $\epsilon = 0.0162$ for Andrés' mechanism under this experiment setting.

When we use the estimated distance error as the location privacy metric, the results are shown in Fig. 3. With this metric, our proposed near-optimal method achieves better privacy than the Andrés' mechanism, but worse than Shokri's optimal obfuscation mechanism (this is reasonable since this mechanism takes the maximum estimated distance error as the optimized object).

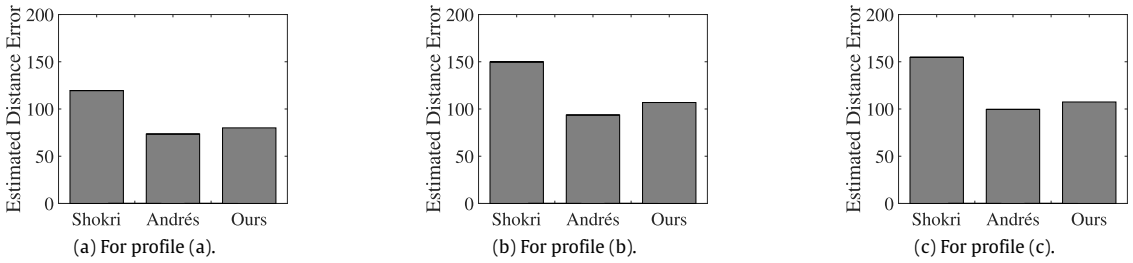
When we use DPLO as the location privacy metric, the results are shown in Fig. 4. With this metric, our proposed near-optimal method achieves much better privacy than the other two mechanisms. This means that with our proposed method, the adversary will obtain the least extra knowledge.

6.3. Experiments on geolife dataset

Using latitude–longitude geographic coordinate system, most user locations are within an area ranges from (116.295E, 39.965N) to (116.355E, 40.015N) in Beijing. We divide the map of this area into 50×50 regions as shown in Fig. 5(a). The latitudinal and longitudinal extent of each region is 0.005, i.e., about 426.6 m $\times$ 556.6 m in size. We focus on the top 20 regions with the highest density, as shown in Fig. 5(b). We treat the set of these 20 regions as $\mathcal{R}$, and compute location privacy with both metrics for all compared methods as shown in Fig. 6.

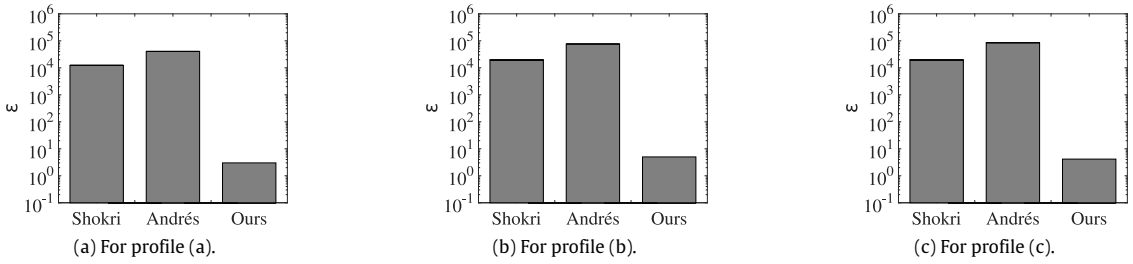
For the estimated distance error, Shokri's mechanism performs the best since it takes the maximum estimated distance error as the optimized object, however this does not always mean better privacy according to our discussion in Section 3.4. Our proposed

[1] In [4], this value is said to be 107.03 m which is mistaken but of no great importance.



(a) For profile (a). (b) For profile (b). (c) For profile (c).

Fig. 3. Estimated distance error (higher is better).



(a) For profile (a). (b) For profile (b). (c) For profile (c).

Fig. 4. ϵ-DPLO (lower is better).

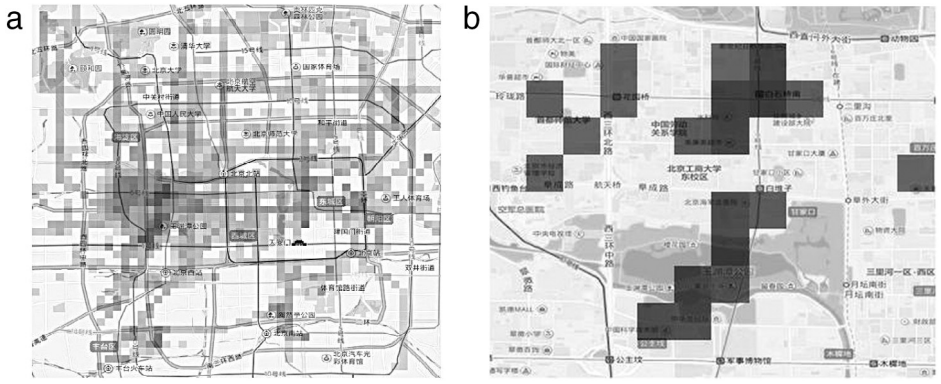
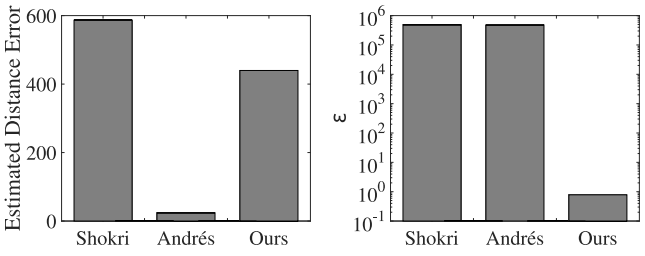


Fig. 5. Geolife user profile in Beijing: (a) Spatial histogram showing the density of users per region in log scale. (b) Top 20 regions with the highest density.



(a) Estimated distance error (higher is better). (b) ϵ-DPLO (lower is better).

Fig. 6. Location privacy.

method performs reasonably and a little bit worse than Shokri's optimal mechanism using this metric. Andrés' mechanism performs much worse than the other two mechanisms, since it is the only mechanism which assumes the user profile to be unknown and makes no profit from this prior knowledge. More over, it seems that Andrés' mechanism performs much worse in real profile (Fig. 6(a)) than in simulated profile (Fig. 3), this contrast mainly comes from the difference between the continuity of the Laplace noise function (also that of the distribution of the simulated profile), and the irregularity and sparsity of real user locations.

For the ϵ-DPLO, our proposed near-optimal method achieves much smaller ϵ than the other compared mechanisms. This means that with our proposed method, the adversary will obtain the least extra knowledge. With our definition, it means better privacy.

7. Discussion and future work

The near-optimal algorithm we use (Algorithm 2, and Algorithm 1) does have high complexity, and it takes about 1–2 h to compute a result for a user's profile with 20 different locations in our experiments. The experiments are performed by 7 python programs running in parallel on a desktop in Ubuntu 14.04, with 7.7 GB memory and 8 i7-4770 CPUs. This algorithm cannot be directly applied on mobile devices to compute an obfuscation in real-time, but it can be performed offline by a powerful server and finally generates a personal privacy rule for any given user profile which performs like a guide on how to hide his location. The privacy rule describes the probability distribution of reporting any obfuscating location with any given real location. With this rule, the obfuscation can be generated on a mobile device in real-time

(within several milliseconds). We believe there are ways to optimize the gradient process by applying some general optimization method like [22]. We leave it as our future work.

8. Conclusion

In this paper, we survey on existing notions of location privacy and make detailed analysis on their limitations. We introduce a new notion of privacy, by quantifying the difference between the prior and posterior knowledge of adversary. With this notion, we show that an optimal obfuscation mechanism can be constructed by solving a non-linear optimization problem. We propose a near-optimal mechanism, and compare it with the state-of-the-art obfuscation mechanisms, using both our proposed metric and the estimated distance error. The results show that under the same quality constraints, our proposed mechanism can achieve better privacy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61320106007, 61402104, 61502100, 61532013, 61572130, 61602111, and 61632008, and by the Jiangsu Provincial Natural Science Foundation of China under Grant BK20140648, BK20150628, and BK20150637, and by Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] M. Gruteser, D. Grunwald, Anonymous usage of location-based services through spatial and temporal cloaking, in: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, ACM, 2003, pp. 31–42.
[2] A.R. Beresford, F. Stajano, Mix zones: User privacy in location-aware services, in: PerCom Workshops, 2004, pp. 127–131.
[3] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, J.-Y. Le Boudec, Protecting location privacy: Optimal strategy against localization attacks, in: Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS'12, ACM, New York, NY, USA, 2012, pp. 617–627. http://dx.doi.org/10.1145/2382196.2382261. URL http://doi.acm.org/10.1145/2382196.2382261.
[4] M.E. Andrés, N.E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, Geo-indistinguishability: Differential privacy for location-based systems, in: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS'13, ACM, New York, NY, USA, 2013, pp. 901–914. http://dx.doi.org/10.1145/2508859.2516735. URL http://doi.acm.org/10.1145/2508859.2516735.
[5] R. Shokri, G. Theodorakopoulos, J.Y.L. Boudec, J.P. Hubaux, Quantifying location privacy, in: 2011 IEEE Symposium on Security and Privacy, 2011, pp. 247–262 http://dx.doi.org/10.1109/SP.2011.18 ISSN: 1081-6011.
[6] C. Dwork, Differential Privacy, in: M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (Eds.), Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1–12. http://dx.doi.org/10.1007/11787006_1.
[7] J. Krumm, A survey of computational location privacy, Pers. Ubiquitous Comput. 13 (6) (2009) 391–399 URL http://dx.doi.org/10.1007/s00779-008-0212-5.
[8] L. Sweeney, k-anonymity: A model for protecting privacy, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10 (05) (2002) 557–570.
[9] C.-Y. Chow, M.F. Mokbel, X. Liu, A peer-to-peer spatial cloaking algorithm for anonymous location-based service, in: Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems, ACM, 2006, pp. 171–178.
[10] B. Hoh, M. Gruteser, H. Xiong, A. Alrabady, Preserving privacy in gps traces via uncertainty-aware path cloaking, in: Proceedings of the 14th ACM Conference on Computer and Communications Security, ACM, 2007, pp. 161–171.
[11] J. Meyerowitz, R. Roy Choudhury, Hiding stars with fireworks: location privacy through camouflage, in: Proceedings of the 15th Annual International Conference on Mobile Computing and Networking, ACM, 2009, pp. 345–356.
[12] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber, Privacy: Theory meets practice on the map, in: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE'08, IEEE Computer Society, Washington, DC, USA, 2008, pp. 277–286 URL http://dx.doi.org/10.1109/ICDE.2008.4497436.
[13] S.-S. Ho, S. Ruan, Differential privacy for location pattern mining, in: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, SPRINGL'11, ACM, New York, NY, USA, 2011, pp. 17–24. http://dx.doi.org/10.1145/2071880.2071884. http://doi.acm.org/10.1145/2071880.2071884.
[14] R. Dewri, Local differential perturbations: Location privacy under approximate knowledge attackers, IEEE Trans. Mob. Comput. 12 (12) (2013) 2360–2372. http://dx.doi.org/10.1109/TMC.2012.208.
[15] N.E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, Optimal geo-indistinguishable mechanisms for location privacy, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS'14, ACM, New York, NY, USA, 2014, pp. 251–262. http://dx.doi.org/10.1145/2660267.2660345. URL http://doi.acm.org/10.1145/2660267.2660345.
[16] M. Li, H. Zhu, Z. Gao, S. Chen, L. Yu, S. Hu, K. Ren, All your location are belong to us: Breaking mobile social networks for automated user location tracking, in: Proceedings of the 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc'14, ACM, New York, NY, USA, 2014, pp. 43–52. http://dx.doi.org/10.1145/2632951.2632953. URL http://doi.acm.org/10.1145/2632951.2632953.
[17] K. Fawaz, K.G. Shin, Location privacy protection for smartphone users, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS'14, ACM, New York, NY, USA, 2014, pp. 239–250. http://dx.doi.org/10.1145/2660267.2660270. URL http://doi.acm.org/10.1145/2660267.2660270.
[18] Y. Xiao, L. Xiong, Protecting locations with differential privacy under temporal correlations, in: Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, CCS'15, ACM, New York, NY, USA, 2015, pp. 1298–1309. http://dx.doi.org/10.1145/2810103.2813640. URL http://doi.acm.org/10.1145/2810103.2813640.
[19] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, J.P. Hubaux, Hiding in the mobile crowd: LocationPrivacy through collaboration, IEEE Trans. Dependable Secure Comput. 11 (3) (2014) 266–279. http://dx.doi.org/10.1109/TDSC.2013.57.
[20] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, ACM Trans. Knowl. Discov. Data (TKDD) 1 (1) (2007) 3.
[21] Y. Zheng, X. Xie, W.-Y. Ma, GeoLife: A collaborative social networking service among user, location and trajectory., IEEE Data Eng. Bull. 33 (2) (2010) 32–39.
[22] Y. Yu, H. Qian, Y.-Q. Hu, Derivative-Free Optimization via Classification, AAAI, 2016, pp. 2286–2292.

Kai Dong is currently an assistant professor at the School of Computer Science and Engineering in Southeast University, China. He received the Ph.D. degree in Computer Science in 2014 from Nanjing University. His research interests include security, privacy, localization and social networks.

Taolin Guo is currently a Ph.D. candidate in Computer Science and Technology in Southeast University, China. He received the Master degree in Software Engineering in 2013 from Southeast University, His research interests include data mining, personal privacy and data security in online social networks and location-based services.

Haibo Ye is currently an Assistant Professor in the College of Computer Science and Technology Nanjing University of Aeronautics and Astronautics, China. He received the Ph.D. degree in computer science in 2016 from Nanjing University. His research interests include indoor localization, mobile and pervasive computing.

Xuansong Li is currently an assistant professor in the School of Computer Science and Engineering at Nanjing University of Science and Technology, China. He received the Ph.D. degree in Computer Science in 2016 from Nanjing University. His research interests include software methodology, formal methods and pervasive computing.

Zhen Ling is currently an assistant professor at the School of Computer Science and Engineering in Southeast University, China. He received the Ph.D. degree in Computer Science in 2014 from Nanjing Institute of Technology and Southeast University, respectively. He joined Department of Computer Science at the City University of Hong Kong from 2008 to 2009 as a research associate, and then joined Department of Computer Science at the University of Victoria from 2011 to 2013 as a visiting scholar. His research interests include network security, privacy, and forensics.