A Framework for measuring query privacy in Location-based Service

Xuejun Zhang[1,2], Xiaolin Gui[1] and Feng Tian[1]
[1] School of Electronic and Information Engineering Xi'an Jiaotong University
Xi'an, Shaanxi 710049 - China
[e-mail: zxjlyl_new@stu.xjtu.edu.cn]
[2] School of Electronic and Information Engineering Lanzhou Jiaotong University
Lanzhou, Gansu 730070 - China
*Corresponding author: Xuejun Zhang

Received December 29, 2014; revised February 11, 2015; accepted March 19, 2015; published May 31, 2015

Abstract

The widespread use of location-based services (LBSs), which allows untrusted service provider to collect large number of user request records, leads to serious privacy concerns. In response to these issues, a number of LBS privacy protection mechanisms (LPPMs) have been recently proposed. However, the evaluation of these LPPMs usually disregards the background knowledge that the adversary may possess about users' contextual information, which runs the risk of wrongly evaluating users' query privacy. In this paper, we address these issues by proposing a generic formal quantification framework,which comprehensively contemplate the various elements that influence the query privacy of users and explicitly states the knowledge that an adversary might have in the context of query privacy. Moreover, a way to model the adversary's attack on query privacy is proposed, which allows us to show the insufficiency of the existing query privacy metrics, e.g., k-anonymity. Thus we propose two new metrics: entropy anonymity and mutual information anonymity. Lastly, we run a set of experiments on datasets generated by network based generator of moving objects proposed by Thomas Brinkhoff. The results show the effectiveness and efficient of our framework to measure the LPPM.

Keywords: location-based services, query privacy, LBS privacy protection mechanisms, privacy metric, measurement

NSFC (Project No. 61472316), Scientific and Technological Project in Shaanxi Province (Project No.2014JQ8322), Basic Science Research Fund in Xi'an Jiaotong University (No.XJJ2014049, No.XKJC2014008), and Shaanxi Science and Technology Innovation Project (2013SZS16-Z01/P01/K01)

http://dx.doi.org/10.3837/tiis.2015.05.009 **ISSN : 1976-7277**

1. Introduction

In recent years, the growing popularity of smart mobile devices equipped with Global Positioning System (GPS) chips, in combination with the increasing availability of wireless data connection, has fostered the development of a variety of successful location-based services (LBSs). Some LBS examples [1] include GPS navigation(e.g., TomTom), mapping applications(e.g., Google Maps), Points of Interest retrieval (e.g., AroundMe), coupon providers(e.g., GroupOn), and location-aware social networks (e.g., Foursquare).

In spite of the enormous benefits brought to individuals and society, LBSs raise a serious privacy concerns as exposure of users' location contained in the LBS queries has been shown to make users susceptible to a broad set of location-based inference attacks, allowing the untrusted/unknown LBS service provider (LSP) to learn private users' information such as their home and work addresses, life styles, political/religious associations, and health conditions. For example, from the anonymous GPS data of individuals it is possible to infer his points of interest (i.e., his home location and work address) [2-4], to predict his past, current and future locations [5-6], and or even to infer his society relationships [7].

In the literature, two major privacy concerns in LBSs have been studied – location privacy and query privacy [8] in terms of the types of sensitive information. The former refers to users' private information directly related to their location containing in a LBS query, as well as the other private information that can be inferred from the location [9]. For example, a user issuing a LBS query in the hospital premise would enable the adversary correlate medical condition to the user. Query privacy, the focus of our paper, refers to users' private information related to LBS query attributes [9]. For instance, the frequent queries for nearest betting office may disclose the user' gambing habits to the adversary. That is, location privacy means hiding the user's location while query privacy means preventing the mapping of a LBS query to a user. The basic idea to protection users' query privacy in LBS is to break the link between user identities and LBS queries [10]. Intuitively, one way to implement this is to anonymize queries by removing or replacing users' identities with pseudonyms. However, this has been proved insufficient, since the adversary can usually find publicly available contextual information (e.g.,white pages) to link the user's identity with her (home) location, and thereby compromise the user's query privacy. In this case, the user's location information can serve as quasi-identifiers. To address the challenge, many research efforts have recently been dedicated to develop LBS Privacy Protection Mechanisms (LPPMs) that allow users to make use of the LBSs while limiting the amount of disclosed sensitive information [8, 11-15]. Most of them are based on intelligently perturbing the information submitted to the LSP, in order to increase the uncertainty of the adversary about user's true location. More specifically, they employ a k-anonymity based framework to blur user exact locations into cloaking regions (CR) so that a certain number of users (at least k) share the same quasi-identifier with the real issuer. Therefore, it is difficult for the untrusted LSP to deduce who is the issuer of the query among the k potential issuers. The caculation of the regions is termed as cloaking or generalisation. As depicted in **Fig. 1**, with this framework, a user issues her location to LSP via a trusted thirty part (TTP) which subsequently strips off the identifier, generates a k-anonymity CR that covers not only query issuer but also k-1 other users geographically. The LSP replies to this CR request and routes the answer through the TTP to be redirected to the specific user with a refined result when possible.

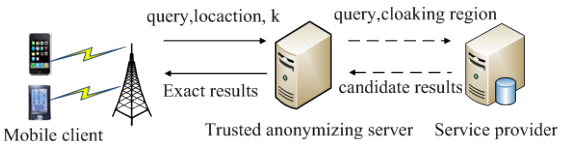

Fig. 1. The trusted thirty server architecture

However, the evaluation of these LPPMs usually neglects that the adversary might have some knowledge about contextual information and also about the algorithm implemented by LPPM. Such information can help the adversary reduce uncertainty on user's true location [16]. Hence, the prior evaluation that disregards such information overestimates the level of privacy protection offered by the LPPM. For example, location k-anonymity is insufficent to measure query privacy when taking into account users' profiles [17]. Nowadays, the popularity of social networks and the growing exposure of peoples' information on Internet provide adversaries source to gather enough background knowledge to obtain the more contextual information [18]. Thus, new privacy risks will emerge. For example, "center-of-ASR" and "outlier" attacks are found on some existing cloaking algorithm when their implementation is made public [19]. Hence, it is a new challenge to quantify the LBS query privacy when the adversary has the knowledge about contexutal information. Furthermore, in current LBS privacy research activities, most efforts focus on developing LPPMs. On the contrary, the methods that evaluate the trustworthiness of the LBSs system, gauge the LBS privacy level of the users, and measure the effectiveness of given LPPM are immature and underdeveloped. Obviously, the lack of a unified and generic formal framework for specifying and evaluating LPPMs is evident, and a good model for the knowledge of the adversary and his possible reasonable ability is also missing. This can lead to a wrong estimation of the query privacy of user.

In this paper, we propose a privacy quantifying framework for modeling and evaluating LBS query privacy regarding the knowledge of contextual information available to the adversary. One of the main ideas of the framework is to explicitly define the assumptions on the knowledge as well as on his reasoning ability. The other elements that influence the user's query privacy, i.e., user's spatiotemporal position, privacy requirements and the LPPMs, are comprehensively studied, especially while various LBSs are being used. Leveraging on this framework, we correctly evaluate the effectiveness of the LPPMs with respect to the knowledge about contextual information, helping user select the appropriate privacy requirements to determine the right tradeoff between privacy and service quality.

2. Related Work

In the LBS privacy protection community, query privacy is the ability to prevent other parties to learn the issuers of queries. So, protection of users' query privacy is essentially to prevent the adversary to learn their issued queries. Tecniques proposed to protect query privacy can be classified into dummy query [14, 20], location cloaking [11-13], and cryptographic transformation [21]. The dummy query methods hide dummy queries among a set of different queries such that the real queries are hidden in the ones. The location cloaking methods exploit the concept of k-anonymity to hide the real issuer in a number of users such that s/he is indistinguishable from the others from the view of the adversary. In the cryptographic transformation methods, users' queries are encrypted and remain secret for LSP so as to offer strong privacy protection. All of these methods introduce extra processing overhead. In this

paper we focus on location cloaking and use it to protect query privay with respect to contextual informaiton. With the aim of assessing the effectiveness of cloaking techiniques, location k-anonymity has been investigated deeply and was first introduced by Gruteser and Grunwald [8] by extending the concept of k-anonymity in database privacy [22]. This metric refers to the situation in which the location precision contained in an LBS query is decreased to a much large area where the query issuer is indistiguishable from at least k-1 other users also present in that area. Because of its simplicity, location k-anonymity has been widely adopted in many different LPPMs, including IntervalCloaking [8], CliqueCloak [11], Casper [23], hilbASR [19], incremental clique-based cloak [13] and dichotomicPoints [17]. However, deeper understanding of k-anonymity reveals its drawbacks. Lin et al. [24] indicate that the k-anonymity is not sufficient enough to reflect the true anonymity when the adversary has the knowledge of different query probability of all users in the cloaking set. Shokri et al. [25] evaluate the effectiveness of k-anonymity in different scenarios in terms of adversary's background information and conclude that location k-anonymity is only effective for protecting query privacy but not location privacy. Subsequently, Shokri et al. [26] propose a distortion-based privacy metric for measuring location privacy, which considers the knowledge of contextual information such as user's mobility patterns, LPPM algorithm and the adversary. Chen et al. [17] indicate that the effectiveness of location cloaking can be compromised when the adversary has access to additional contextual information (e.g., user profiles) which have many interpretations in the literature. Shokri et al. [16] use user's mobility profile and propose a more general use of the adversary's expected estimation error to quantify location privacy, taking into account the adversary's knowledge of user mobility pattern, LBS access pattern, and the internal algorithm implemented by LPPM. Howerver, they do not consider the knowledge of user's profiles. Personal information (e.g., gender, job, salary) is usually available on the Internet, i.e., online social network, and can serve as user profiles as well. Shin et al. [27, 28] propose k-anonymity based metrics by restricting levels of similarity among users in CR in terms of their profiles.

Based on above research, we can find: (1) the existing query privacy metrics are not sufficent for LBS when considering the adversary's background knowledge as well as his reasoning ability, and (2) most of the existing query privacy metrics are designed for specific LPPMs. The lack of a unified and generic formal framework for the evaluation of the LBS query privacy is evidence.

In this paper, our main goal is to define a common formal framework for measuring query privacy in LBSs. We focus on evaluation of the query privacy with regards to an individual query rather than query histories. Moreover, we make use of users' static and public personal information such as profession, gender, age, preference as user's profiles. Considering the users' query histories and mobile pattern is part of our future works.

3. Quantification framework of Query privacy

In this section, we present our framework for query privacy. This allows us to precisely specify its relevant components and attacks on query privacy with the contextual information. In this paper, we only consider location cloaking as in LBSs users require instant response. To properly assess their effectiveness under the different adversary models, we comprehensively consider the elements that influence the query privacy of users, including their spatiotemporal positions, privacy and service quality requirements, LPPM, adversary model and privacy metrics etc. As these elements are highly interconnected, they should be studied together in a consistent framework. We define such a framework (as shown in **Fig. 2**) as a tuple of the

following inseparable elements: <***U***, Q, **L**PPM, $\hat{Q}$, **A**ttacker, **M**etric>, where ***U*** is the set of mobile users who move within an area and subscribe the LBS queries whose geographical and contextual information is embodied in location contained in the *Q*, and *Q* represents the set of queries issued by the user. LBS query privacy protection mechanism, **L**PPM, distorts the query q (a member of *Q***)** and produces the generalization query $\hat{q}$ (a member of $\hat{Q}$, which is the set of observable queries to an adversary). Hence, when accessing the LBS, users only expose the output of **L**PPM, instead of sharing their actual queries. The **A**ttacker is an entity who implements inference attacks to infer some information about *Q* (e.g., issuer and location of a given user at a given time instant) having observed $\hat{q}$ and by relying on his knowledge about contextual information. The evaluation metric, **M**etric, captures the performance of the adversary and his success in re-identifying the expected information about queries. Note that inference attack of adversary is in the sense of statistics. The adversary utilizes the compromised contextual information (e.g., user's profiles) to extract the priori probability knowledge $p(q_j|u_i)$, forming the probability matrix $\boldsymbol{M}=(m_{ij})$, where element $m_{ij} = p(q_j|u_i)$. We use $M(q_i)$ to denote the *i-th* row of ***M***, the probability distribution over users to issue the query q_i. The objective of adversary is to exploit probability matrix ***M*** to reconstruct posterior probability of each user in the *CR* to be issuer when query q_i is observed.

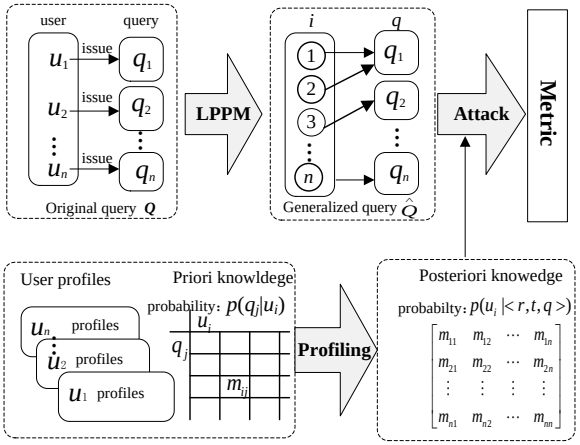

Fig. 2. Query privacy quantifying framework

In the following subsections, we present and specify all the entities and components of our framework and illustrate their inter-relationship.

3.1 Mobile Users and LBS Queries

In mobile wireless network environment, mobile users usually use their location-aware wireless devices to connect to LBS through the wireless infrastructure (e.g., cellular and WiFi networks) in a certain spatiotemporal status. As users move between locations, they leverage the infrastructure to submit original local-based queries to an LSP via a TTP which transforms original queries into the generalized ones, at some frequency. To formally define the original query and generalized query, we use $\boldsymbol{U}=\{u_1,u_2,\cdots,u_N\}$ to denote a set of *N* mobile users who move within an area that is partitioned into *M* different regions $\boldsymbol{L}=\{l_1,l_2,\cdots,l_M\}$. Time is also considered to be discrete, and the set of time instances that can be recorded is $\boldsymbol{T}=\{1,2,3,\cdots,\mathcal{T}\}$. The granularity of time instances is determined by LBS. Let ***whereis***: $\boldsymbol{U}\times\boldsymbol{T}\rightarrow\boldsymbol{L}$ to be a

function that gives the actual location of a user at any time instance. The user spatial distribution at time t can be defined as the set $\boldsymbol{S}(t)=\{(u, \boldsymbol{whereis}(u,t))| u\in \boldsymbol{U}\}$.

Definition 1 (Original Query) Let $\mathcal{Q}$ be the set of queries supported by LBS, e.g., nearest gas station. We say that a quadruples $< u, t, \boldsymbol{whereis}(u,t), q_c>$ is a original query, where $u\in \boldsymbol{U}$ is the identities of users, $t\in \boldsymbol{T}$ represents time stamp when the query is issued, $q_c\in \mathcal{Q}$ is the query contents. The set of queries from users $\boldsymbol{U}$ at a given time $\boldsymbol{T}$ is denote by $Q\subseteq \boldsymbol{U}\times \boldsymbol{T}\times \boldsymbol{L}\times \mathcal{Q}$. Note that an element $q=<u,t,\boldsymbol{whereis}(u,t),q_c>$of Q represents the actual status of users in the realistic mobile network environment.

Definition 2 (Generalised query) Let $\mathcal{Q}$ be the set of queries supported by LBS, $\boldsymbol{I}$ be the set of all pseudonyms used by all users, $\boldsymbol{P}(\boldsymbol{L})$ is the power set of $\boldsymbol{L}$, and then we use $\boldsymbol{R}\subset \boldsymbol{P}(\boldsymbol{L})$ to denote the set of all possible CRs. We say that a quadruples $<i,t,r,q_c>$ is a generalised query, where $i\in \boldsymbol{I}$ is user's pseudonyms and i may be null that represents the status of being identity-less (when a user's identity is removed from his communication messages without being replaced by one of his pseudonyms), $r\in \boldsymbol{R}$ and $\boldsymbol{whereis}(i,t)\in r$, $q_c\in \mathcal{Q}$. The set of generalized queries, corresponding to the Q, is denoted by $\hat{Q}\subseteq \boldsymbol{I}\times \boldsymbol{T}\times \boldsymbol{R}\times \mathcal{Q}$. Note that an element $\hat{q}=<i,t,r,q_c>$ of $\hat{Q}$ represents the spatiotemporal status of users from the perspective of observer.

3.2 LBS Privacy Protection Mechanisms

Mobile users share their original queries (e.g., Q) with possibly untrusted LSP in various LBSs, or unwilling reveal their identities (e.g., U) and locations (e.g., L) contained in the original queries to curious eavesdropping adversaries through wireless channel. In all these scenarios, an adversary can track or identify users over an observation period, unless their identities and locations are properly modified before being exposed to others. The mechanism that performs this modification in order to protect the users' query privacy is referred to a LPPM.

As mentioned above, we focus on cloaking-based LPPM in which the TTP use cloaking algorithm to transform an original query into a generalized query and forwards it to the LSP. The main idea is to cloak the data that could be used, possibly joined with external knowledge, to re-identify individuals. More specifically, given an original query $q=<u,t,\boldsymbol{whereis}(u,t),q_c>$, the LPPM replaces the user's identity (i.e., u) with the pseudonym (i.e., i) and substitute a large area (i.e., r) for user's location (i.e., l) to protect query privacy. Formally, we represent the LPPM as a function: $f: Q\rightarrow \hat{Q}$. For instance, we have $f(q)=\hat{q}$. The function f maps the original queries Q to generalised query $\hat{Q}$, which also takes users' privacy requirements as part of its input. The objective of adversary is to invert this mapping by depending on their knowledge: Given the subset of generalised queries, he tries to re-identify the real issuer of the original queries.

3.3 The Adversary

In the LBS system, privacy risks and countermeasures should be categorised according to the adversary's model and goals [10]. For query privacy, the adversary's goal is to associate issuers to their queries while the model should be defined in terms of his knowledge and attack(s) [28]. As the adversarial knowledge and attack(s) has directly influence on the difficulty of attacking this LPPM. The more knowledge to the adversary has available and the more his reasoning abilities the higher the probability of inferring the users' private

information. Prior works [19, 25] have indicated that the effectiveness of a LPPM can only be formally evaluated only if the reidentify capabilities of the adversary are explicitly stated. To define such capabilities, we introduce the knowledge of the adversary as a contextual information, denoted C, of a possible attack that consists of a set of assumptions. In this paper, we assume that some contextual information that the adversary has access to is inherently contained in C.

A. The assumption on adversarial knowledge and abilities

To integrate the adversarial knowledge and reasoning capabilities in the quantification of query privacy, we need to know how much or what knowledge is available to adversary. In reality, this might not be infeasible. Consequently, it is common to make some assumptions on adversarial knowledge and capabilities. We assume an adversary: (1) intercepts all generalised queries forwarded by TTP. This assumption implies that either the LSP is untrusted or the communication channel between the TTP and the LSP is not secure; (2) knows the LPPM used by TTP (i.e., f). This assumption is common in the literature since data security techiques are typically public; (3) knows all users' current spatial distribution (i.e., $S(t)$). This assumption is conservative but possible. In real scenarios, users may often issue LBS query from the same positions (home, office). Referring to the address books or other public information, the real issuer can be identified through physical observation, triangulation, etc. In the worst case, an adversary may be able to obtain the positions of all users in the anonymity set of the query. This represents a very strong adversary which allow us to analyse query privacy in the worst case. The availability of $S(t)$ make the adversary to obtain the set of users presented in any region r at tim t, which is denoted as $AS_C(r,t)$; On the other hand, because in practice the adversary does not have all users' locations, it is important that the cloaking algorithm does not reveal the position of any user, to avoid giving away additional information. (4) has no knowledge about the decision process of users' privacy requirements. However, an adversary can learn the users' privacy requirements after observing the generalised queries. This is realistic. As from the features of generalised queries, the adversary can infer the corresponding privacy requirements; (5) cannot link any two queries from the same user. All queries are independent from the adversary's perspective; (6) exploit users' profiles to obtain the prior knowledge over users regarding the issuing queries.

Once the contextual assumptions are defined, we can formalize how the adversary can try to infer, from a generalized query, the real issuer that issued it. Given a generalised query $\hat{q}$ and C, we model this attack as the likelihood of associating a specific identity to a generalized query $\hat{q}$ from the view of the adversary with C. For any user $u \in U$, the corresponding probability distribution can be respented as $p(I=u|\hat{q},C)$, where I is the issuer of $\hat{q}$. We refer to this as a posterior probability of user u, which can be computed as follows:

$$p(I=u\mid\hat{q},C)=\frac{p(u,\hat{q},C)}{p(\hat{q},C)}=\frac{p(C)\cdot p(u\mid C)\cdot p(\hat{q}\mid u,C)}{\sum_{u'}p(C)\cdot p(u'\mid C)\cdot p(\hat{q}\mid u',C)}$$
$$=\frac{p(u\mid C)\cdot p(\hat{q}\mid u,C)}{\sum_{u'}p(u'\mid C)\cdot p(\hat{q}\mid u',C)}$$

In the above equation, the distribution $p(u|C)$ represents the probability that user u issue a query at time t based on the contextual informaiton C. Since there has no information about the distribution available, we can assume it as uniform according to the principle of maximun entropy [29]. For $\forall u' \in u$, this leads to $p(q\mid C)=p(u'\mid C)$. Thus, the posterior distribution can be simplified as:

$$p(I=u\mid \hat{q},C)=\frac{p(\hat{q}\mid u,C)}{\sum_{u'\in U} p(\hat{q}\mid u',C)} \tag{1}$$

The probability $p(\hat{q}\mid u,C)$ indicates the likelihood that if user *u* generates a original query q at time *t* then the q will be generalized as $\hat{q}$. This is actually a joint of the following two probabilities: 1) the probability that user *u* issues the original query q when he submit a query at time *t*. We call this probability the a priori probability of user *u*; 2) the probability that the LPPM takes as input q and outputs $\hat{q}$. We use $p(q|u,C)$ and $p(f(q)=\hat{q})$ to denote these two probabilities, respectively. Thus, we have

$$p(\hat{q}\mid u,C)=p(q\mid u,C)\cdot p(f(q)=\hat{q}) \tag{2}$$

We assume that the LPPM mentioned in this paper are deterministic. That is, there is always a unique generalized query corresponding to each original query. This implies that the $p(f(q)=\hat{q})$ is either 0 or 1. Given an original query and a generalised query, this value is availabe to the adversary because LPPMs are public. Therefore, the key of query privacy analysis is to compute $p(q|u,C)$ for any query $q\in Q$.

The calculation of $p(q|u,C)$ depends on *C*, i.e., the available contextual information. In this paper, we only consider the adversary's knowledge about static contextual information, i.e., users' profiles. In other words, the contextual information does not change over time. Note that in practice we can also consider it as static if a type of contextual information keep stable for a sufficiently long period. For example, a user's profiles can be considered as static even though the user may change his job as changing jobs is not frequent. In the following discussion, we give the method of computing a priori knowledge based on contextual information *C*.

B. The derivation of the adversarial knowledge based on user's profiles

User profiles are associated with a set of attributes that characterize the user. These attributes may contain the description information (e.g., age, job, gender, nationality), contact information (e.g., zip codes, name, address, e-mail) and personal preferences (e.g., hobbies, favourite activities, moving pattern) [27]. The values of these attributes can be categorical (e.g., nationality) or numerical (e.g., salary, age), and can be discretized into a categorical or interval form. For example, the value of a home address can be represented by the corresponding zone in which it lies while the numerical values of age can be discretised into three intervals, such as '≤20', '≥20,≤40', '≥40'. Note that the intervals are mutually exclusive and their union is equal to the original domain. In this way, each attribute has a finite number of candidate values.

Let $A=\{a_1,a_2,\cdots,a_n\}$ be the list of the attributes where a_i is the name of the attribute. Each attribute has its certain domain. The profiles of user *u* can be represented as $\varphi_u=\{a_1: val_1, a_2: val_2,\cdots, a_n: val_n\}$, where val_i is the corresponding value of attribute a_i, denoted by $\varphi_u^{a_i}$. Because not all the attributes are applicable to all users, some of them may be empty for certain users. Thus the contextual information learnt by the adversary can be represented as the following:

$$C=\{S(t), f, \{\varphi_u \mid u\in U\}\}.$$

Our main idea of calculating the $p(q|u,C)$ is to compute the relevance of user *u*'s profile to each query and compare the relevance to *q* with those to other queries. Given a profile attribute

a_i, we can discretize its domain into intervals if it is numberical or divide the domain into sub-categories if it is categorical. After being discretized, each profile attribute can be denoted by binary digits. We simply use the bits as much as the number of all the possible discrete values for a_i. With this binary string of bits, we can denote the profiles of user u as a vector $P_u^{a_i}=[l_1,l_2,\cdots,l_n]$, where l_i is sequence of binary digits of discrete values of a_i. The length of l_i is equal to the number of all possible discrete values of a_i, and the digital is 1 if val_i satisfies the corresponding discrete value and 0 otherwise. For instance, since all the possible values of a profile attribute 'gender' are "female" and "male", we utilize the two bits | M | F | to represent the gender: "female" can be indicated with '01'and "male" can be indicated with '10'. Moreover, the numerical attribute 'age' can be discretized into the form of | ≤20 | ≥20,≤40 | ≥40 |. Thus, the age 25 can be represented as '010'.

Each query $q\in Q$ should have a set of correlated attributes that can be used to deduce the real issuer of this query q. Furthermore, for a given related profile attribute, its value has different contribution to identify the real issuer. Therefore, each value of a relevant attribute has a different weight to measure the probability that the user issues the given query q. For instance, for the query asking for expensive luxury, the associated attributes should include job, salary and age while nationality is irrelevant. Among them, a salary is much more relevant than age and moreover, a salary of more than 10 000 dollar is much more important than one of less than 1000 dollar. Hence, we use a relevant vector $W_q^{a_i}$ for each attribute a_i to express the relation between values of attributes and queries. Let $W_q^{a_i}=[w_1,w_2,\cdots,w_n]$ be the relevance vector of query q of attribute a_i. For any $u\in \boldsymbol{U}$ and $q\in Q$, let $\mu(u,q)=\sum_{i\le n} w_q^{ai}\cdot \boldsymbol{P}_u^{a_i}$ be the relevance value of user u's profile to query q. Consequently, the probability of user u issuing the query q based on contextual information C is:

$$p(q\,|\,u,C)=\frac{\mu(u,q)}{\sum_{q'\in Q}\mu(u,q')} \tag{3}$$

3.4 Privacy metrics

In the literature, a large number of coaking-based LPPMs have been proposed to protect query privacy by departing the association between users' identities and their queries. The objective of the adversary is to try to invert these LPPMs, depending on his observed generalized query and contextual informaion C. The performance of the adversary and his success in recovering the desired information about $\hat{q}$ captures the level of the query privacy offered by these LPPMs. Therefore, the performance of the adversary need to be quantified precisely in order to improve the performance of LPPMs. Furthermore, LBS users need the measurement to express their privacy requirements for their queries.

Beside location k-anonymity, numerous privacy metrics have been proposed for quantifying the capability of the adversary, such as feeling-based [30], expected estimation error-based [25, 31]. The feeling-based metric employs location entropy to quantify the average uncertainty of the adversary to guess the issuer in a given scenario. The estimation error-based metrics quantify privacy as the probability of the adversary choosing the real issuer when he makes a single guess. The probabilistic nature of the adversary's task implies that he, in most of the cases, cannot be completely sure of the issuer. Uncertainty is thus inevitable. We use a posterior probability distribution to capture the adversary's certainty and

quantify the expected correctness of his attack. In this section, we present three metrics on query privacy and formally define them using our framework.

Definition 3 (**location *k*-anonymity**) Let $q=<u,t,\textbf{\textit{whereis}}(u,t),q_c>\in Q$ to be the original query, $\hat{q}=<i,r,t,q_c>\in \hat{\boldsymbol{Q}}$ be the corresponding generalized query. The issuer u is k-anonymity if $|\{u\in \boldsymbol{U}|\textbf{\textit{whereis}}(u,t) \in r \wedge \boldsymbol{f}(q)=\hat{q}\}|\geq k$.

The definition 3 shows that all users in the anonymity set are all k-anonymity as they take r as the generalised region for the query $\hat{q}$ at time t. Location k-anonymity quantify privacy as the ability of the adversary to differentiate the real issuer from the other k-1 users within the anonymity set $AS_C(r,t)$.

However, as discussed in the section 2, location k-anonymity is not sufficient for measuring users' query privacy when users' profiles are regarded as the part of the adversary's knowledge. Particularly, the user whose posterior probability is higher than others is easy to be selected as the issuer candidate.

In this paper, the attacker's objective is to explore the observed query $\hat{q}$ and the knowledge about user's contextual information C to infer the real issuer who is in the anonymity set $AS_C(r,t)$. The higher the uncertainty of the user's identity associated with the query is, the harder the adversary infers the real issuer. Location entropy is a widely used metric for measuring the uncertainty associated with location information in LBS queries. Thus, location entropy can also be used to describe the adversary's certainty to identify the issuer of a generalized query in our context. The higher the entropy is, the lower the adversary's certainty is. Let variable $\boldsymbol{I}$ denotes the issuer of a generalized query $\hat{q}$. Then the adversary's certainty can be expressed as equation (4):

$$E(I \mid \hat{q}, C) = -\sum_{u\in AS_C(r,t)} p(u \mid \hat{q}, C)\cdot \log p(u \mid \hat{q}, C) \tag{4}$$

From the perspective of the adversary, users located in the generalized region r have the same possibilities to issue their queries when the adversary has no knowledge about users' contextual C. At this point, the entropy is maximum. So the adversary's certainty is minimum. In contrast, certain users located in the region r are more likely to be taken as the candidates for the issuer from the adversary's view than others also in this region if the adversary gains more knowledge about certain users' contextual information. Thus the entropy will be minimum.

For a given generalized query $\hat{q}$ and a given β values, we say that the issuer is β entropy anonymity if all users in anonymity set $AS_C(r,t)$ can have r as their generalized regions when issuing the same query, and the entropy $E(I \mid \hat{q}, C)\geq\beta$.

Definition 4 (**β entropy anonymity**) Let $\beta>0$, $q=<u,t,\textbf{\textit{whereis}}(u,t),q_c>\in Q$ is the original query, $\hat{q}=<i,r,t,q_c>\in \hat{\boldsymbol{Q}}$ is the corresponding generalized query. The user u is β entropy anonymity if

$$E(I\mid \hat{q}, C)\geq\beta \wedge \boldsymbol{f}(q)=\hat{q}$$

With this metric, users can specify appropriate entropy values to express their privacy requirements. The users' privacy requirements consist of a metric and its corresponding value, such as (β entropy anonymity, 2.8).

For a given LPPM, the amount of the knowledge about contextual information gained by the adversary has directly influence on the difficulty of attacking this LPPM. The more information gains the adversary obtains, the higher the probability of identification of real issuer is. Mutual information is a useful tool in information theory, which quantifies the

mutual dependence of two random variables. Thus we can use mutual information to evaluate the certainty increased after revealing the generalized query. Before the adversary compromises the generalized query, he only knows the u's priori probabilit $p(q \mid u, C)$. Thus, the certainty of the attacker can be described as $E(Q \mid u, C)$. After the adversary compromise a generalized query $\hat{q}$, his certainty can be expressed as $E(I \mid \hat{q}, C)$. Therefore, for a given query q, the amount of information certainty gained by the adversary after observing the corresponding generalized query $\hat{q}$ can be expressed as equation (5):

$$\begin{aligned} M(I \mid q;\hat{q}) &= E(Q \mid u, C) - E(I \mid \hat{q}, C) \\ &= -\sum_{q \in Q} p(q \mid u, C) \cdot \log p(q \mid u, C) \\ &+ \sum_{u \in AS_C(r,t)} p(u \mid \hat{q}, C) \cdot \log p(u \mid \hat{q}, C) \end{aligned} \tag{5}$$

The value of $M(I \mid q;\hat{q})$ reflects the changes of certainty to the adversary. The greater the $M(I \mid q;\hat{q})$ is, the lower the adversary's certainty is, and vice versa.

For a given generalized query $\hat{q}$ and a given value δ, we say that the issuer u is δ mutual information anonymity if all users in anonymity set $AS_C(r,t)$ get the same generalized area r when issuing query q and $M(I \mid q;\hat{q}) \leq \delta$.

Definition 5 (**δ mutual information anonymity**) Let δ>0, q=<u,t,***whereis***(u,t),q_c>$\in Q$ is an original query, $\hat{q}$=<i,r,t,q_c>$\in \hat{\boldsymbol{Q}}$ is the corresponding generalized query. The issuer u is δ mutual information anonymity if $\forall u \in AS_C(r,t)$ satisfy

$$M(I \mid q;\hat{q}) \leq \delta \text{^} \boldsymbol{f}(q) = \hat{q}.$$

With this metric, users can specify the appropriate value of mutual information to express their privacy requirements, such as (δ mutual information anonymously, 4.0).

4. Experiment analysis

In this section, we use our framework to measure the effectiveness of the query privacy protection algorithms "DichotomicPoints" [17] which consider the adversary's knowledge and ability and can defense against the "outliers" attack. The cloaking algorithm DichotomicPoints takes as input the original query q, user's privacy requirements (location k-anonymity, β entropy anonymity, δ mutual information anonymity and their corresponding values), the user's real-time spatial distribution S(t) and prior probability matrix $\boldsymbol{M}$. the output is the generalized area r which satisfies the user's privacy requirement.

To generate the coordinates of mobile users and their LBS queries, we use a network based generator of moving objects proposed and implemented by Brinkhoff [32]. We randomly generate 10 000 mobile users and simulate their movement on the real road map of Oldenburg (region area of 23.57km × 26.92km), a city in Germany. For the moving speeds, we use the default setting in the generator, which changes users' speeds at each intersection based on the road type. **Fig. 3** shows the global view of map of Oldenburg with footprints of mobile users.

The adversary's knowledge about users' profiles can be obtained by using Adult dataset in UCI machine learning and methods proposed in section 3.3. As we focus on assessing our framework, we randomly generate users' priori probability in the experiments. We implemented the algorithm DichotomicPoints using Java and the simulation experiments are

run on a Win7 PC with 3.51 GHz Intel Core (TM) i5 processor and 4GB memory. The results are obtained by taking the average of 100 times simulation of the corresponding algorithms.

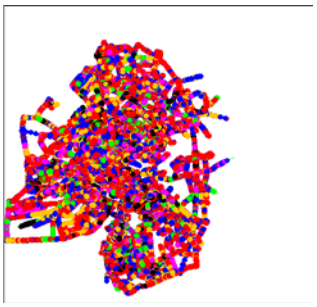

Fig. 3. Oldenburg datasets

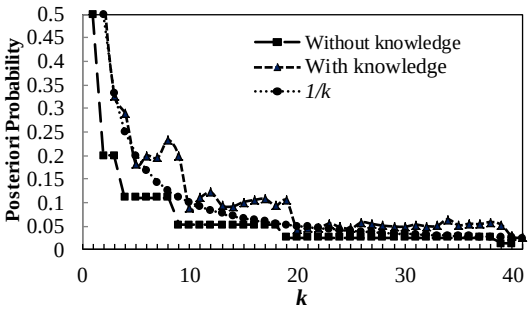

Fig. 4. Location k-anonymity metric

A. k-anonymity metric

We evalute the effectiveness of our framework by checking if it can increase the likelihood of the adversary to correctly identify real issuers of generalze query by obtaining contextual information. Given a generalised query, we can use the issuer's posterior probability as the measurement of the correctness of the adversary's attack on query privacy [16]. If the contextual information can assist in compromising user's query privacy, then issuer will have larger posterior probability than those computed without the information on average. In **Fig. 4**, we show how the average a posteriori probability of the issuer changes with k regarding that the adversary has the knowledge about user's profiles or not. Without the knowledge about user's profiles, the adversary only knows that all the users in the generalized area r have the same possibility to issue the generalised query. In other words, the adversary can only infer that the probability of each user issuing query $\hat{q}$ is the reciprocal of the number of users in anonymity set $AS_C(r,t)$ (corresponding to the curve of "*without knowledge*"). With the knowledge about users' profiles, the attacker knows that the probability of each user issuing query is not the same and would select the user whose posterior probability is the largest as the real issuer of $\hat{q}$ (corresponding to the curve of "*with knowledge*").

In **Fig. 4**, we observe that the user's posterior probability decreases as k increases. This is because large k means more users are in the generalized region r. Give a k, the issuer's posterior probability is normally less than $1/k$ if the adversary has no knowledge about users' profiles. That is to say, the adversary cannot identify the real issuer of query with the probability larger than $1/k$. The size of k can correctly reflect the privacy level of user. On the contrary, when there have more contextual information (e.g., user's profiles) available to the adversary, the issuer's posterior probability is normally large than $1/k$. That is, the adversary can confidently identify the real issuer if the probablity is sigificantly high for the real issuer. Thus the k-anonymity metric is not sufficent for measuring users' query privacy.

From the above dicussion, we can conclude that our framework is useful to increase the likelihood of adversary to correctly identify the real issuer. Consequently, it is necessary to integrate the adversary's background knowledge and ability in the privacy qualification. In addition, the results of privacy evaluation should include the query privacy level of user and the assumptions on knowledge available to the adversary.

B. β entropy anonymity and δ mutual information anonymity metrics

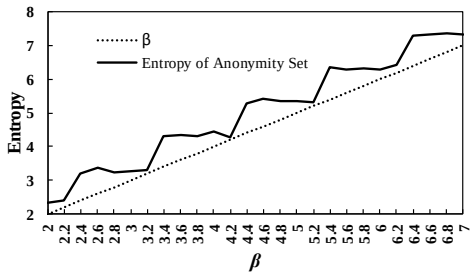

Fig. 5. β entropy anonymity

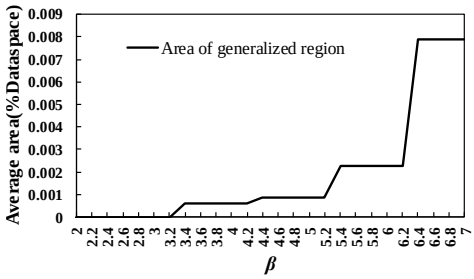

Fig. 6. Area of β entropy anonymity regions

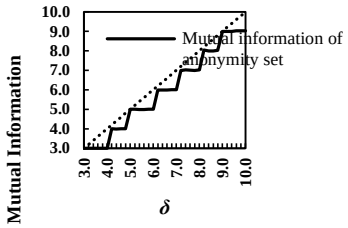

Fig. 7. δ mutual information anonymity

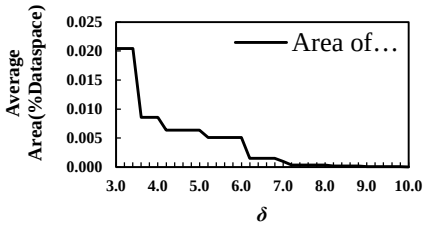

Fig. 8. Area of mutual information anonymity regions

In **Fig. 5** and **Fig. 7**, we show that the entropies and mutual information corresponding to the generalised regions generated by the dichotomicPoints algorithm satisfy the definitions of β entropy anonymity and δ mutual information anonymity. We can observe that the values of entropy (resp. mutual information) change sharply when β (resp. δ) is getting close to integers. This is due to the nature of entropy. **Fig. 6** and **Fig. 8** show that the average area of generalised regions changes along with β and δ respectively. In **Fig. 6**, the average area usually increases as β increases. This is because the larger the β is, the greater the entropy over users in the generalized region is and the more users the generalized region contains. Therefore, the privacy level of the user in the generalized region is higher, but the service quality is lower. Similarly, in **Fig. 8**, the area decrease as δ increases. This is because the larger the δ, the less the attacker's uncertainty is and the less users the generalized region contains. Therefore, the privacy level of the user in the generalized region is lower, but the service quality is higher.

From the above dicussion, we can see that the β entropy anonymity metric and δ mutual information privacy can correctly reflect the privacy level of user provided by the LPPM with respect to the knowledge available to the adverary and can help user determine the tradeoff between the privacy requirements and service quality requirements.

5. Conclusion

This paper presents a formal framework for specifying LBS query privacy exploring contextual information. The framework is to comprehensively take the various elements and relations together that influence the query privacy of users into account and to formally define assumption on the knowledge and ability available to the adversary. Moreover, one way to model the adversary's attack on query privacy is proposed in the framework. The privacy metrics are also described in the framework. Experiment results demonstrate the effectiveness

of our framework to evaluate query privacy. Simultaneously, it also shows that it is necessary to contemplate the attacker's knowledge and ability in the measurement of LPPMs. Furthermore, it is need to consistently model users' requirements together with the adversary's knowledge and objective for designing a new query privacy protection mechanism.

As a follow-up to this work, we will incorporate the location-based applications into the framework and analyse the effectiveness of query privacy protection mechanisms with respect to these applications. Furthermore, we will focus on the model of users' mobility patterns and query history in the framework.

Acknowledgements

The work was partially supported by grants from NSFC (Project No. 61172090, 61163009, and 61163010), National Science and Technology Major Project (Project No. 2012ZX03002001), Ph.D. Programs Foundation of Ministry of Education of China (Project No. 20120201110013), Scientific and Technological Project in Shaanxi Province (Project No. 2012K06-30).

References

[1] M. E. Andrés, N. E. Bordenabe, "Geo-indistinguishability: Differential privacy for location-based system," in *Proc. of the 20th ACM Conf. on Computer and Communications Security*, pp. 901-914, 2013. Article (CrossRef Link)

[2] J. Krumm, "Inference attacks on location tracks," *in Pervasive Computing*, vol. 4480, pp. 127-143, 2007. Article (CrossRef Link)

[3] J. Freudiger, R. Shokri, J.-P.Hubaux, "Evaluating the privacy risk of location-based services" in *Proc. of the 15th Int'l Con. on Financial Cryptography and Data Security*, pp. 31-46, 2011. . Article (CrossRef Link)

[4] S. Gambs, M.-O. Killijian, M. N. Prado, "Show me how you move and I will tell you who you are," *Transactions on Data Privacy*, vol. 2, no, 4, pp. 103-126, 2011. . Article (CrossRef Link)

[5] M. C. Gonzalez, C. A. Hidlgo, Blbert-Laszlo, "Understanding individual human mobility pattern," *Nature* vol .453, pp. 779-782, 2008. Article (CrossRef Link)

[6] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no 5968, pp. 1018-1021, 2010.

[7] S. Gambs, M.- O. Killijian, "De-anonymization attack on geolocated data," in *12th IEEE Int'l Conf. on Trust, Security and privacy in Computing and Communications*, pp. 789-797, 2013. Article (CrossRef Link)

[8] M. Gruteser, D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. of the First Int'l Conf. on Mobile Systems, Application, and Services*, pp. 31-42, 2003. Article (CrossRef Link)

[9] K. G. Shin, X. E. Ju, Z. G. Chen, and X. Hu, "Privacy protection for users of location-based services," *IEEE Wireless Communications*, vol. 12, no. 1, pp.30-39, 2012. Article (CrossRef Link)

[10] C. Bettini, S. Mascetti, X. S. Wang, D. Freni and S. Jajodia, "Anonymity and historical k-anonymity in location-based services," in *Privacy in Location-Based Applications*, vol. 5599, pp. 1-30, 2009. Article (CrossRef Link)

[11] B. Gedik, L. Liu, "Protecting location privacy with personalized k-anonymity: architecture and algorithms," *IEEE Transaction on Mobile Computing*, vol. 7, no. 1, pp. 1-18, 2008. Article (CrossRef Link)

[12] T. Xu, Y. Cai, "Exploring historical location data for anonymity preservation in location-based services," in *Proc. of the 27th IEEE Int'l Con. on Computer Communications*, pp. 1220-1228,

2008. Article (CrossRef Link)
[13] X. Pan, J. L, Xu, X. F. Meng, "Protection Location Privacy against Location-Dependent Attacks in Mobile Services," *IEEE Transaction on knowledge and data engineering*, vol. 24, no. 8, pp. 1506-1519, 2012. Article (CrossRef Link)
[14] A. Pingley, N. Zhang, X.W. Fu, H-A Choi, S. Subramaniam, W. Zhao, "Protection of query privacy for continuous location based services," in *Proc. of the 30th IEEE Int'l Con. on Computer Communications* , pp. 1710-1718, 2011. Article (CrossRef Link)
[15] R. Shokri, G .Theodorakopoulos, C. Troncoso, et al, "Protecting Location Privacy: Optimal Strategy against Localization Attacks," in *Proc. of the 19th ACM Conf. on Computer and Communication Security*, pp. 617-626, 2012. Article (CrossRef Link)
[16] R. Shokri, G. Theodorakopoulos, J-Y L. Boudec et al, "Quantifying location privacy," in *Proc of the 32nd IEEE Symposium on Security and Privacy*, pp. 247-262, 2011. Article (CrossRef Link)
[17] X. H. Chen, J. Pang, "Measuring Query Privacy in Location-Based Services," in *Proc. of the second ACM Conf. on Data and Application Security and Privacy*, pp. 49-60, 2012. Article (CrossRef Link)
[18] X. Wu, X, Ying, K. Liu, and L. Chen, "A survey of algorithms for privacy-preservation of graphs and social networks," *in Managing and Mining Graph Data*, pp. 421-442, 2010. Article (CrossRef Link)
[19] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries," *IEEE Transaction on Knowledge and Data Engineering*, vol. 19, no. 12, pp. 1719-1733, 2007. Article (CrossRef Link)
[20] H. Kido, Y. yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proc. IEEE. Int'l Conf. Pervasive Services*, pp. 88-97, 2005. Article (CrossRef Link)
[21] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.–L. Tan, "Private queries in location based services: anonymizers are not necessary," in *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pp. 121-132, 2008. Article (CrossRef Link)
[22] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, vol. 10, no. 5, pp. 557-570, 2002. Article (CrossRef Link)
[23] M. F. Mokbel, C.Y. Chow et al, "The new Casper: query processing for location services without compromising privacy," in *Proc. of the 32nd Int'l Conf. on Very large data bases*, pp. 763-774, 2006. . Article (CrossRef Link)
[24] Lin X, LI S P, YANG Z H, "Attacking algorithms against continuous queries in LBS and anonymity measurement," *Journal of Software*, vol. 20, no. 4, pp. 1058-1068, 2009. Article (CrossRef Link)
[25] R. Shokri, C.Troncoso, C. Diaz, "Unraveling an Old Cloak: k-anonymity for Location Privacy," in *Proc. of the 2010 ACM Workshop on Privacy in the Electronic Society*, pp. 115-118, 2010. Article (CrossRef Link)
[26] R. Shokri, J. Freudiger, M. Jadliwala, J.-P. Hubaux, "A distortion-based metric for location privacy," in *Proc. of the 8th ACM Workshop on Privacy in the Electronic Society*, pp. 21-30, 2009. Article (CrossRef Link)
[27] H. Shin, V. Atluri, J. Vaidya, "A profile anonymization model for privacy in a personalized location base service environment," in *Proc. of the 9th Int'l Conf. on Mobile Data Management*, pp. 73-80, 2008. Article (CrossRef Link)
[28] H. Shin, V. Atluri, and J. Vaidya. "A profile anonymization model for location-based services," *Journal of Computer Security*, vol. 19, no. 5, pp.795-833, 2011. Article (CrossRef Link)
[29] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review Series Ⅱ*, vol. 106, no. 4, pp. 620-630, 1957. Article (CrossRef Link)
[30] T. Xu, Y. Cai, "Feeling-based location privacy protection for location-based services," in *Proc. of 16th ACM Conf. on Computer and Communication Security*, pp. 348-357, 2009. Article (CrossRef Link)

[31] X. J. Zhang, X. L. Gui, F. Tian, S. Yu, J. An, "Privacy quantification model based on Bayes conditional risk in location-based services," *Tsinghua Science and Technology*, vol. 15, No. 5, pp. 452-462, 2014. Article (CrossRef Link)
[32] T. Brinkhoff, "A framework for generating network-based moving objects," *GeoInformatica*, vol. 6, no. 2, pp. 153-180, 2002. Article (CrossRef Link)

Xuejun Zhang is a PhD candidate student in Department of Computer Science and Technology at Xi'an Jiaotong University, China. He was also a visiting scholar at Wuhan Unversiting, China, during 2010 to 2011. He got his M.S. degree from Southeast University, China, in 2008. His research interests include location privacy protection and quantification, and service computing and security.

Xiaolin Gui is currently a professor, PhD supervision and Deputy Dean of school of Electronic and Information Engineer, Xi'an Jiaotong University, China. His research covers secure computation of open network system including Grid, P2P, and cloud computing, dynamic trust management theory, data and privacy protection, and Internet of things. He got his PhD, MEng and BEng degrees from Xi'an Jiaotong University in 2001, 1993 and 1988, respectively. He has published over 130 academic papers and books.

Feng Tian is a PhD candidate student in Department of Computer Science and Technology at Xi'an Jiaotong University. He got his BEng degree from Xi'an Jiaotong University in 2009. His research interests include location privacy protection and data outsourcing security.