PAPER Special Section on Security, Privacy and Anonymity of Internet of Things

Privacy-Aware Information Sharing in Location-Based Services: Attacks and Defense

Zhikai XU[†a)], ***Student Member***, **Hongli ZHANG**[†b)], **Xiangzhan YU**[†c)], ***Nonmembers***, ***and*** **Shen SU**[†d)], ***Student Member***

SUMMARY Location-based services (LBSs) are useful for many applications in internet of things(IoT). However, LBSs has raised serious concerns about users' location privacy. In this paper, we propose a new location privacy attack in LBSs called hidden location inference attack, in which the adversary infers users' hidden locations based on the users' check-in histories. We discover three factors that influence individual check-in behaviors: geographic information, human mobility patterns and user preferences. We first separately evaluate the effects of each of these three factors on users' check-in behaviors. Next, we propose a novel algorithm that integrates the above heterogeneous factors and captures the probability of hidden location privacy leakage. Then, we design a novel privacy alert framework to warn users when their sharing behavior does not match their sharing rules. Finally, we use our experimental results to demonstrate the validity and practicality of the proposed strategy.

key words: *location privacy, inference attack, privacy enhancing technology, location-based services, internet of things*

1. Introduction

With the rapid development of GPS-enabled mobile devices and wireless communication technology, location-based services (LBSs) have been one of the novel uses and most popular activities in internet of things (IoT). LBSs offer several attractive functions, such as getting information about nearby Points of Interest (POIs), locating nearby friends, and sharing geo-content with friends, etc. However, a user's location history can be analyzed to infer much more information about the user than the locations themselves, which can lead to unwanted information leakage. The ability to analyze data may lead to privacy mining from simple location information, particularly in the era of big data. The potential abuse of users' location information by unauthorized entities is evolving into a serious concern.

Some may argue that users are increasingly aware of the value, potential, and risks of publishing geo-location content. Users should be aware of the issue of privacy leakage and specify the connected users or location-enabled applications that can access users' location information. However, although users are typically able to choose their information sharing behavior, they are often unable to verify whether their sharing selections match their preferred information-sharing policies for the following reasons: 1) users want to release as much geo-location content as possible to receive better services; and 2) users lack knowledge about risks in the era of big data. Even privacy-aware users may make uninformed decisions. Sometimes they are not even aware of the leakage of private location information. Thus, it is not sensible to place the burden of verifying location privacy leaks on users.

In recent years, many approaches [1]–[6] have been proposed to address the location privacy protection problems in LBSs, including spatial cloaking [2], [3], dummy locations [4], [5], and anonymity [6], [7]. However, in the era of big data, these approaches are vulnerable to *side information* based inference attacks, in which the information multiplies and is shared ever more widely. An adversary can easily extract a great deal of side information from various types of communications, which are summarized as follows:

(1) Geographical information, e.g., the road network and the distance between POIs.

(2) Check-in history, e.g., users' check-in information that is available via the internet.

(3) User preferences, e.g., users' ratings of POIs and information in their homepages, such as interests and historical check-in POIs.

This paper aims to address the privacy-aware location information sharing problems in LBSs. The key challenges include accurately and efficiently evaluating whether sharing behavior meets users' sharing rules and designing a privacy alert framework for LBSs users. In response to these challenges, we first evaluate and model the impact of these factors - spatial context information, human mobility patterns, and user preferences - and their separate impacts on users' check-in behaviors. Then, we propose a novel algorithm based on Multiple Linear Regression to integrate the above-referenced heterogeneous factors and to provide a personalized check-in behavior prediction for each user. Finally, a novel privacy alert framework for LBSs users is proposed to warn users when their actual sharing behaviors are not congruent with their sharing rules. Therefore, this paper makes the following contributions:

1) We discover and model three key factors that can be used to infer a user's private location information: geographic information, human mobility patterns and user preferences.

Manuscript received November 26, 2015.
Manuscript revised April 6, 2016.
Manuscript publicized May 31, 2016.

[†]The authors are with the School of Computer Science Engineering, Harbin Institute of Technology, Harbin, 150001, China.
a) E-mail: zhikaixu@foxmail.com
b) E-mail: zhanghongli@hit.edu.cn
c) E-mail: yxz@hit.edu.cn
d) E-mail: johnsuhit@gmail.com

DOI: 10.1587/transinf.2015INP0001

2) We develop a novel algorithm to optimally integrate the multiple heterogeneous factors discussed above into a single framework that outperforms state-of-the-art methods.

3) We design a novel privacy alert framework to warn users when their sharing behavior does not correspond with their sharing rules.

4) We conduct extensive experiments to evaluate the performance of our methods using two real-world datasets.

The remainder of this paper is organized as follows. Section 2 discusses related work on location privacy. Section 3 introduces our motivation and basic definitions. Our proposed algorithms are presented in Sect. 4. Section 5 proposes the privacy alert framework. Section 6 presents the experimental results, which is followed by concluding remarks in Sect. 7.

2. Related Work

Recently, protecting the location privacy of mobile users has become a popular topic in IoT. Most current solutions adopt the k-anonymity model [1], which stipulates that location information contained in a message sent from a mobile user to a location-based service provider(LBSP) should be indistinguishable from at least k-1 other messages from different users, where k is a user-specified anonymity requirement. Depending on the different methods that are employed to distort the user's query before it reaches the LBSP, the previous research can be classified into the following categories:

(i) Spatial cloaking [2], [3]. The basic objective of spatial cloaking is to blur users' exact locations into spatial regions to meet users' privacy requirements, such as k-anonymity, l-diversity, or t-closeness. Then, instead of their exact locations, users send spatial regions to the LBS. However, in the era of big data, this method is vulnerable to *side information*-based inference attacks.

(ii) Dummy location [4], [5]. Users protect their location privacy by reporting their real location together with many faked locations - also known as dummies - to the LBSP. However, this method does not apply to any network that emphasizes content authenticity. Moreover, how to generate dummy locations is still an open question.

(iii) Anonymity or pseudonym [6], [7]. These types of approaches attempt to protect users' location privacy by removing their real identities or replacing them with pseudonyms. In [6], Barkhuus et al. found that mere anonymity does not protect users' privacy and has proposed that users change pseudonyms over time in a mix zone, which refers to a small area, such as a road intersection. Unfortunately, this mix-zone approach is not suited for a real-name network.

(iv) Encryption [8], [9]. In this method, users' locations are encrypted and queries are processed using ciphertexts. For example, in [8], Zhao et al. proposed a location privacy preserving framework based on searchable encryption (SE). In [9], Li et al. proposed a fine-grained privacy preserving location query protocol based on holomorphic encryption (HE). The drawback to this approach is that its cryptographic computations are costly for mobile users.

In addition, most of the above approaches focus on the "single shot" scenario and fail to take into account the *side information* beyond the user's maximum velocity. However, in the era of big data, such information multiplies and is shared worldwide. The adversary can easily extract substantial knowledge about a user from various types of information.

Recently, some approaches have been proposed to address the inference attack based on *side information*. Andres et al. in [10] proposed a notion of geo-indistinguishability which extends differential privacy and guarantees privacy against "any side information". However, a fundamental difference is that it achieves geo-indistinguishability by location perturbation and it does not consider the temporal correlations of multiple locations. T. Murakami et al. in [11] quantified the risk of de-anonymization in a realistic situation where users disclose only a small amount of location information to the attacker. It proposed a method to overcome a data sparsity problem when the attacker trains a transition matrix that is specific to each user. Assuming that the adversary knows the query probability in each location, Niu et al. in [4] proposed an entropy-based dummy location selection algorithm, in which the dummy locations with similar query probabilities are preferentially chosen.In [12], Xue et al. proposed predicting a trip's destination by human mobility patterns and presented an algorithm that can minimize the number of locations a user must hide to avoid privacy leakage. In [13], Sadilek et al. utilized a dynamic Bayesian-based inference model, using users' social ties to infer their locations. In [14], Huo et al. proposed an inference model based on a Hidden Markov Model (HMM), using similar users' check-in histories to predict the unobserved locations.

However, the studies discussed above only use mobility pattern [12] or social information [13], [14] to predict users' behaviors. In addition, the prior preference in these studies for different factors is not personalized for users. In contrast to the above approaches, our method has the following features: (i) we integrate three heterogeneous factors to capture the probability of hidden location privacy leakage; (ii) we provide a personalized behavior prediction for each user; and (iii) we propose a privacy alert framework to warn users when their sharing behavior does not correspond with their respective sharing rules.

3. Preliminaries

In this section, we first present our motivation and then formally define the secure metrics.

3.1 Motivation

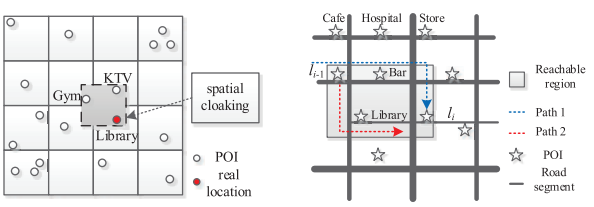
Big data has become a hot topic in many fields, including market research, targeted advertising, workflow improvement, and national security. It is considered a revolutionary technology that will transform how we live, work, and think. On the other hand, in the era of big data, protecting

location privacy is becoming more difficult as information multiplies and is shared even more widely throughout the world. An adversary with big data analytical capacity can easily extract a great deal of *side information* from diverse mass data and then use it to infer user's privacy. Exist location protection approaches always fail to consider the side information in adversaries' hands. Due to the ignorance on the *side information*, there schemes cannot effectively guarantee the desired privacy protection level. We will illustrate the *side information* based inference attack with two examples.

Example 1: Alice wants to find her nearby friends but she does not want to give away her exact location. Thus, as shown in Fig. 1(a), instead of the exact location, Alice blurs her real location into spatial regions to meet her anonymity constraints (such as k-anonymity) and reports the spatial region to her friends. The above approach, known as "spatial location cloaking"[2,3], is a classic method used in the era of location privacy. However, it is vulnerable to *side information*-based inference attacks. An adversary that can access Alice's check-in history may find that she enjoys reading (e.g.,Alice's published spatial regions often contain bookstores), thus inferring that she is more likely to be in the library than in other places. To achieve the desired privacy protection level, Alice should use larger k or cloak more wider area.

Example 2: Bob does not want anyone to know he has been to the hospital. Therefore, as shown in Fig. 1(b), he disables location services to avoid privacy leakage while at the hospital, whereas he releases his location at the cafe and store. In the era of big data, however, this strategy does not necessarily prevent an adversary from inferring that the user has visited the hospital. An adversary with big data mining ability may gain the knowledge that most users will follow a path cafe->hospital->store and compute the travel time between the cafe and the store. The adversary may infer that Bob visited the hospital because his travel time between the cafe and the store is longer than the average interval. To achieve the desired privacy protection level, Bob should hide wider area. That is, if the path information related with the road below the hospital is totally hidden, the attack could be protected.

Though the countermeasures in example 1 and 2 are secure enough by tuning the parameters of the countermeasures. Excessive protection may result in serious quality of service degradation perceived by LBS users. In addition, because LBS users lack knowledge about the risks(e.g., all the users' check-in history), they cannot logically compare their actual sharing behavior with their preferred privacy protection level; sometimes, they are not even aware of location privacy leakage.



(a) Spatial cloaking:3 anonymity (b) Adaptive information share

Fig. 1 Side information based inference attack

Therefore, this paper is motivated to address the privacy-aware location information sharing problem in LBSs. In this paper, *side-information* is limited to the information obtained from the knowledge from check-in sequences and open street map. We don't assume any other *side-information* (e.g., information deduced from twitter or blog). We propose a privacy evaluation system under the assumption above. By using our system, we can detect the vulnerability of countermeasures shown in example 1,2, and tune the better parameters of countermeasures for preventing the vulnerability. Our goal is not to prevent users from sharing location information but to let users share as much location information as possible when their privacy requirements are satisfied. To achieve this goal, we propose a novel privacy metric that formally represents different users' personal privacy requirements, and we then explore approaches that meet those requirements.

3.2 Threat Model

In this paper, we make a widely acceptable assumption that the location-based services provider (LBSP) in IoT provides honest but curious services. LBSP acts in an honest fashion and correctly follows the designated protocol specification. However, it is interesting in analyzing users' published locations so as to infer the users' sensitive information.

3.3 Basic Definitions

In this paper, we take check-in service as an example for all the discussions and illustrations. However, it can be easily extended to other LBSs in IoT. The following are the definitions used in the article.

Definition 1. Share rules: In LBSs, a user always specifies a group of users or applications that either can or cannot access their location information. These specifications are known as share rules. Three examples are shown below.

Share rule r_1: I do not want anyone know I have been to the hospital.

Share rule r_2: Only my classmates can know my location when I am at the pub.

Share rule r_3: Google Map can not access my location when I am at home.

Definition 2. Hidden location: Given user u_1's share rules, if user u_2 does not have access to u_1's location when u_1 is at POI l_i, then l_i is u_1's hidden location for u_2.

Definition 3. Hidden location inference attack: In the era of big data, the adversary can extract substantial side information about users from various sources, including geographic information, social information, and check-in histories. Based on this side information, the adversary can

infer users' likely hidden locations using their visible locations.

Definition 4 Users' check-in sequence: User u_k's published locations can be represented as a check-in sequence $C(u_k) = (< l_1, t_1 >, \cdots, < l_i, t_i >, \cdots, < l_m, t_m >)$, in which t_i denotes u_k's check-in time at POI l_i. All POIs are sorted by check-in time.

Definition 5 s-confidence privacy: Given the sharing rule r_i and the corresponding hidden location l_i, we say the system provides s-confidence privacy for sharing rule r_i if no adversary can use $C(u_k)$ to deduce that the probability that u_k has visited l_i is higher than s.

Definition 6 Users' privacy requirement: Given user u_k, we represent his privacy requirement as $PR(k) =< (r_1, l_1, s_1), \cdots, (r_i, l_i, s_i), \cdots, (r_n, l_n, s_n) >$, where l_i denotes the hidden POI deduced from u_k's sharing rule r_i, and s_i denotes the privacy requirement for rule r_i (POI l_i) is s_i-confidence. Notably, different users may have different personal sharing rules, and the privacy requirements for various sharing rules may be different.

4. Hidden Location Inference Model

In this section, we first provide an overview of the inference model and then perform an in-depth analysis regarding the impact of different factors on users' check-in behaviors. Finally, we show how different features are fused together to infer users' hidden locations.

4.1 Overview

Through the analysis of users' check-in data, we find that users pay the most attention to the following three aspects when they choose the POIs that they will visit:

1) Geographic information: Typically, the distance to the user significantly affects the probability that the user will the visit the POI. For example, most users will likely choose a nearby place when Foursquare recommends restaurants.

2) Human mobility patterns: Users' behaviors typically follow certain mobility patterns in specific regions. For example, people who check-in at the railway station are then likely to go either to the airport or to another railway station.

3) User preferences: In real life, people like to visit certain POIs that fit their own personal interests. For example, most people who frequent the library also like visiting bookstores.

As shown in Fig. 2, we evaluate the impact of these factors separately from users' check-in behavior. Moreover, we propose a novel multi-factor model based on Multiple Linear Regression to fuse the factors. Finally, we combine the multi-factor model with the current location in the prediction of the POI that the user will visit or has visited. Details are provided in the next subsection.

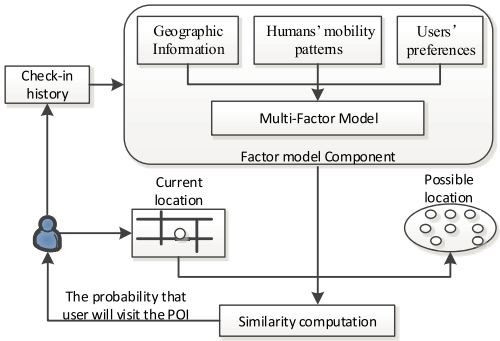


Fig. 2 The overview of the inference model

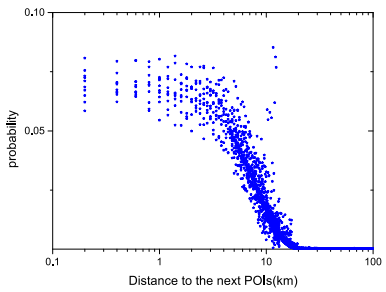


Fig. 3 The probability distribution of distance between the two consecutive check-in POIs

4.2 Factor Modeling

4.2.1 Geographic Information

As we know, the distance between POIs significantly affects users' visitation behaviors. To observe the distance factor in depth, we conduct an analysis using Foursquare data, which was made available by Gao [15]. For each user, we calculate the distance between all the consecutive check-in POIs in his check-in sequence. Figure 3 depicts the probability distribution of distance between the two consecutive check-in POIs. We find that the probability that a user will visit a POI is inversely proportional to the distance between the POI and his/her current location. The correlation between the POIs decreases as the distance between the POIs increases. Thus, we adopt the Gauss formula to calculate the distance similarity between POI l_i and l_j.

$$p_{geo}(l_i, l_j) = exp(-\frac{Distance(l_i, l_j)^2}{2}) \quad (1)$$

where $Distance\ (l_i,\ l_j)$ denotes the Euclidean distance between l_i and l_j.

Moreover, we note that human behavior shows strong daily and weekly periodic patterns, which is not surprising and has been demonstrated by numerous researchers. Thus, in addition to the user's current location, the user's check-in history can also be used to predict POIs that the user will visit next. Given user u_k's check-in history $C(u_k) = \{l_1, \cdots, l_i, \cdots, l_m\}$ and current location l_α, the probability that user u_k will visit POI l_β after l_α can be calculated

as

$$p_{geo}(l_\beta|l_\alpha, u_k) = a p_{geo}(l_\alpha, l_\beta) + (1-a) \sum_{l_i \in C(u_k)} p_{geo}(l_i, l_\beta) \tag{2}$$

where a is turning parameter ranging within [0, 1], and it can be determined via cross-validation.

4.2.2 Human Mobility Patterns

User behaviors typically follow certain mobility patterns in a specific region [16]. For example, people who check-in at the railway station are likely to go either to the airport or to another railway station. Moreover, in some cases, the correlation between the POIs is unidirectional. For example, people usually check-in at gyms before they check-in at restaurants. However, few people are tempted to adopt the reverse order because it is not healthy to exercise right after a meal.

Mobility patterns have important effects on users' check-in behaviors. Thus, a basic approach is to use the majority of users' historical check-in sequences to capture their mobility patterns and then use those mobility patterns to infer their future check-in behaviors. Given all the users' check-in sequences, we can compute the transition probability between POIs, which is defined as

$$p_{hm}(l_j|l_i) = \frac{count(l_i, l_j)}{count(l_i)} \tag{3}$$

where count (l_i, l_j) denotes the number of check-in sequences that contains POI l_i, POI l_j in sequence, and count(l_i) denotes the number of check-in sequences that contains POI l_i. Note that the transition probability is unidirectional and $p_{hm}(l_j|l_i)$ is not equal to $p_{hm}(l_i|l_j)$ in most cases. Then, given user u_k's current location l_α, the probability that u_k will visit POI l_β after l_α is calculated as

$$p_{hm}(l_\beta|l_\alpha, u_k) = p_{hm}(l_\beta|l_\alpha) \tag{4}$$

4.2.3 Users' Preference

In real life, people actually like to visit the POIs that fit their own personal interests. Thus, an adversary can infer the probability that a user will visit a POI by analyzing that user's interests. An example is illustrated in Fig. 1(b). The user disables the location service while at the library but checks in at POIs l_{i-1} and l_i. Because the interval between the check-in behaviors is greater than the regular travel time, the adversary can easily infer that the user has visited another POI between POI l_{i-1} and l_i. Moreover, suppose that by analyzing check-in history, the adversary knows user u_k often visits libraries. The adversary will then be able to infer that user u_k has visited the library instead of the bar, even though most people drive along path 1 between l_{i-1} and l_i.

Definition 4. (Users' Interest Matrix) Given the users' ratings on POIs, we can build a user's interest matrix $I_{|U|\times|L|}$, which is represented as

$$I_{|U|\times|L|} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \dots & r_{mn} \end{pmatrix} \tag{5}$$

where r_{ki} denotes u_k's rating on POI l_i, reflecting how much user u_k likes POI l_i. In the era of big data, it is not difficult for an adversary to obtain a huge number of users' direct or implicit ratings of POIs from various websites, such as Foursquare, Yelp, and Dianping. Moreover, we can calculate the implicit rating r_{ki} by using the number of repeat visits to POI l_i from u_k. Formally, r_{ki} is calculated as

$$r_{ki} = \frac{|q_{ki}|}{\sum_{l_j \in C(u_k)} |q_{kj}|} \tag{6}$$

where q_{ki} denotes the number of repeat visits to POI l_i from u_k, and $C(u_k)$ denotes user u_k's check-in sequence. However, we find this method insufficient. For example, u_k may check-in 10 times at Wal-Mart, where a large number of users will check-in, whereas u_k checks-in 10 times at the museum, where few people go. Equation 6 illustrates that u_k's rating for Wal-Mart is equal to its rating for the museum. However, it is apparent that the museum is more important when analyzing that user's personal preferences.

Inspired by the TF-IDF [17] (term frequency-inverse document frequency) scheme in information retrieval systems, we treat a check-in record as "a term" and treat the user's check-in sequence as "a document". Formally, the user's rating r_{ki} is computed as

$$r_{ki} = \frac{|q_{ki}|}{\sum_{l_j \in C(u_k)} |q_{kj}|} * log \frac{|U|}{|\{u_s | q_{si} \geq 1, u_s \in U\}|} \tag{7}$$

where U denotes the collection of all the users and q_{ki} denotes the number of repeat visits to POI l_i from u_k. The first part of (7) is the TF value of POI l_i in the check-in sequence of u_k, and the second part denotes the IDF value of POI l_i. However, the interest matrix obtained by the above methods is a sparse-matrix. Thus, we first calculate the POI similarity among POIs and then integrate it into a collaborative filtering (CF) algorithm [18] to predict users' ratings on unvisited POIs.

Definition 5. POI similarity $sim(l_i, l_j)$**:** POI correlation indicates the correlation between POI l_i and POI l_j in the space of human behavior. For example, if most people who often go to the library also like visiting the bookstore, then the library and the bookstore are similar in the space of human behavior.

Figure 4 illustrates the computation process: users' ratings on the POIs are stored as entries in the interest matrix where the matrix rows correspond to the users and the columns correspond to the POIs. The basic idea behind calculating the correlation between POI l_i and l_j is to isolate those users who have rated both of the POIs and apply a cosine-similarity to calculate the correlation between the POIs by using only the ratings of the co-rated users. In this case, the two POIs l_i and l_j are considered as two vectors in

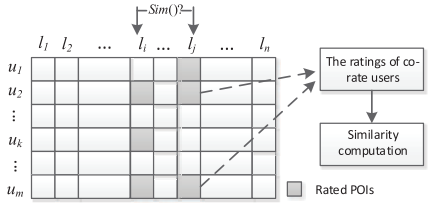


Fig. 4 The process of correlation computation

the m-dimensional user-space, and the correlation between the POIs is measured by computing the similarity between the vectors. Formally, $sim(l_i, l_j)$ is calculated as

$$sim(l_i, l_j) = \frac{\sum_{u_k \in co(l_i,l_j)} r_{ki} \cdot \sum_{u_k \in co(l_i,l_j)} r_{kj}}{|\sum_{u_k \in co(l_i,l_j)} r_{ki}|^2 |\sum_{u_k \in co(l_i,l_j)} r_{kj}|^2} \quad (8)$$

where $co(l_i, l_j)$ denotes the users who have rated both POI l_i and POI l_j, and "$\cdot$" denotes the dot-product between the two vectors. Then, we use the POI similarity and user's rating of visited POIs to predict the user's rating of unvisited POIs. Intuitively, predicting u_k's rating of POI l_j given the user's ratings of POI l_{i-1} and l_i, if POI l_i is more related to l_j beyond l_{i-1}, then u_k's rating of POI l_i is likely to be a far better predictor of POI l_j than u_k's rating of POI l_{i-1}. Formally, our approach can be represented as

$$r_{kj} = \frac{\sum_{l_i \in C(u_k)} (r_{ki} + dev_{l_j,l_i}) * sim(l_i, l_j)}{|\sum_{l_i \in C(u_k)} sim(l_i, l_j)|} \quad (9)$$

$$dev_{l_j,l_i} = \frac{\sum_{u_s \in co(l_i,l_j)} (r_{sj} - r_{si})}{|co(l_i, l_j)|} \quad (10)$$

where $sim(l_i, l_j)$ denotes the similarity between POI l_i and POI l_j, and dev_{l_j,l_i} is calculated using Eq. (10). Then, given user u_k's current location l_α, the probability that u_k will visit POI l_β after l_α is calculated as

$$p_{up}(l_\beta | l_\alpha, u_k) = \frac{r_{k\beta}}{max_{l_i \in C(u_k)}(r_{ki})} \quad (11)$$

where $max_{l_i \in C(u_k)}(r_{ki})$ is a normalization term.

4.2.4 Factor Fusion and Probability Prediction

As different features have different impacts on different users, the challenge is determining how to evaluate the significance of each and then fuse them together to predict the probability that the user will visit the POI. As a challenge, we propose a novel behavior predictor model based on Multiple Linear Regression, which describes the relationship between the user's check-in behavior and the above features.

Multiple Linear Regression [19] is a well-known method used to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data, and it has been used successfully in many different fields, including medicine and economics. The basic idea of Multiple Linear Regression is applicable to our system because it delivers good results and is easy to fine-tune and customize.

As discussed above, in LBSs, check-in behavior is affected by the spatial factor, the sequence factor, and the preference factor. Then, the probability in Eq. (2) is combined with the probabilities in Eqs. (4) and (11) to produce a unified measure through the sum rules:

$$\begin{aligned} p_{total}(l_\beta | l_\alpha, u) = {} & \beta_0 + \beta_1 * p_{geo}(l_\beta | l_\alpha, u) \\ & + \beta_2 * p_{hm}(l_\beta | l_\alpha, u) \\ & + \beta_3 * p_{up}(l_\beta | l_\alpha, u) \end{aligned} \quad (12)$$

where $\beta_0, \beta_1, \beta_2$ and β_3 denote the weights of different features. Given user u_k's check-in history $C(u_k)$, the user's check-in behavior is binary (i.e., 0 or 1), which means she may or may not visit a POI after a specific POI. Formally, given n check-in records, the Multiple Linear Regression model is calculated as follows:

$$Y = X\beta \quad (13)$$

$$Y = (y_1, y_2, \ldots, y_n)^T \quad (14)$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \quad (15)$$

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T \quad (16)$$

where y_i denotes the i^{th} observation, x_{i1} denotes the geographic factor in the i^{th} observation, x_{i2} denotes the human mobility pattern factor in the i^{th} observation, and x_{i3} denotes the preference factor in the i^{th} observation. Given the observations, we can compute the values of $\beta_0, \beta_1, \beta_2$ and β_3 using the least-squares model [20], in which the weight of different features minimizes the sum of the squares of the errors made in the results of every single observation.

The inference model can then integrate these three factors to provide a personalized check-in behavior prediction for each user. Let $p_{nor}(l_\beta | l_\alpha, u_k)$ denote the normalized probability that u_k will visit l_β after l_α, and let L_{next} denote the set of the candidate locations that the user may visit after l_α (the details of the computation of L_{next} will be given in the next section). Then, given any POI l_α, the probability that user u_k will visit l_β after l_α can be calculated as follows:

$$p_{nor}(l_\beta | l_\alpha, u_k) = \frac{p_{total}(l_\beta | l_\alpha, u_k)}{\sum_{l_i \in L_{next}} p_{total}(l_i | l_\alpha, u_k)} \quad (17)$$

where $p_{total}(l_i | l_\alpha, u_k)$ can be calculated using (12). Obviously, $\sum_{l_i \in L_{next}} p_{nor}(l_i | l_\alpha, u_k) = 1$.

5. PLIS: Privacy-Aware Location Information Share Framework

In this section, we first introduce the system architecture of our privacy-aware location share framework. Next, we present the privacy evaluation algorithm.

5.1 System Architecture

Figure 5 depicts the architecture of our privacy-aware location information share (PLIS) framework, which includes three critical components: mobile users, the inference server (IS), and the location-based service provider (LBSP). We describe the details of these components below.

(1) Users: Users in the system are equipped with GPS-enabled smart phones, which are capable of communicating with the IS and the LBSP. To receive service from the LBSP, the user will send his/her location l_i, the current time t_i, and the identities of those friends who can access location f to the IS. The check-in request is represented as $u(l_i, t_i, f)$.

(2) The IS: The IS consists of three main components: a hidden location finder, a probability calculation component, and a privacy estimating component. When the IS receives the request message, it will evaluate whether the user's privacy requirement (sharing rules) will be satisfied; if so, the current location is released to the specified friends f. If there is no location privacy leakage, the user's check-in request $u(l_i, t_i, f)$ will be forwarded to the LBSP automatically; otherwise, the IS will return the sharing rule that has been violated, the probable leaked hidden POIs, and the leakage probabilities of hidden POIs to the user. Then, the user will decide whether to release the current location to his/her friends.

(3) The LBSP: The LBSP provides various services for the users, such as enabling check-in, locating nearby friends, or distributing coupons from businesses close to the user's current location. In addition, the LBSP provides the IS with the side information.

5.2 Algorithm

Take user u_k as an example: Given user u_k's privacy requirement $PR(k)$, the IS can obtain the set of u_k' hidden sensitive POIs $LS(k)$ and his privacy requirement for the POIs. The IS will evaluate whether the user's privacy requirement will be satisfied if the current location is released to the specified friends/applications f when it receives u_k's check-in request $u_k(l_i, t_i, f)$,

First, the IS will evaluate whether the hidden sensitive locations that the user has visited will be leaked if the current location is published. Given u_k's previous check-in record $u_k(l_{i-1}, t_{i-1}, f)$, the IS will calculate the interval between two check-ins: $\triangle t = (t_i - t_{i-1})$. In the era of big data, map information and maximum velocity of the road segments are publicly available via the Internet, and the IS can thus use the side information to evaluate whether u_k has visited other POIs between the two check-ins. If $\triangle t \leq (dis(l_{i-1}, l_i)/v_{max})$, the IS may infer that u_k definitely does not have time to visit any other POIs ($dis(l_{i-1}, l_i)$ denotes the distance between l_{i-1} and l_i). Otherwise, as shown in Fig. 1(b), the IS will calculate the reachable POI set $L_{pre} = \{l_m | dis(l_{i-1}, l_m) + dis(l_m, l_i) \leq \triangle t * v_{max}\}$. Then, for each $l_j (l_j \in (L_{pre} \cap LS(k))$, the IS will check whether the deduced probability that u_k has visited l_j is higher that his privacy requirement for lj if the current location is published. The probability that the user has visited l_j between the two check-ins can be formalized as a posterior probability $p_{nor}(l_j | l_{i-1}, l_i, u_k)$, which can be derived using Bayesian reasoning. Formally, it can be calculated as follows:

$$p_{nor}(l_j | l_{i-1}, l_i, u_k) = \frac{p_{nor}(l_{i-1}, l_j, l_i, u_k)}{p_{nor}(l_{i-1}, l_i, u_k)} = \frac{p_{nor}(l_j | l_{i-1}, u_k) p_{nor}(l_i | l_j, u_k)}{\sum_{l_s \in L_{pre}} p_{nor}(l_s | l_{i-1}, u_k) p_{nor}(l_i | l_s, u_k)} \quad (18)$$

where $p_{nor}(l_s | l_i, u_k)$ is calculated using (17).

Then, the IS will evaluate whether the location that user plans to visit will be leaked if the current location is published. Identifying all the POIs in the city as the candidate locations that the users will visit is unrealistic and useless since the probability that the users will visit POIs that are far or unpopular is low. Therefore, we only take three types of POIs into consideration: (i) the top k POIs nearest to the current location; (ii) the top k POIs that other users will visit; and (iii) u_k's top k highest rated POIs. From this, we can narrow down the set of POIs, L_{next}, that the user may visit. Then, for each $l_j (l_j \in (L_{next} \cap LS(k))$, the IS will check whether the deduced probability that u_k will visit l_j is higher than his privacy requirement for l_j if the current location is published. The probability that the user will visit l_j can be calculated using (17).

If there is no privacy leakage, the IS will forward the user's check-in request $u_k(l_i, t_i, f)$ to the LBSP; otherwise, the IS will return the sharing rule being violated, the probable leaked hidden POIs, and the leakage probabilities of hidden POIs to the user. Then, the user will decide whether to release the location. The details of this are shown in Algorithm 1.

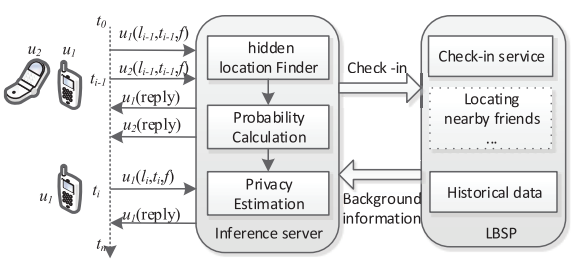


Fig. 5 System architecture of the privacy-aware location information share framework

5.3 Security Analysis

Since cryptography techniques such as SHA and AES can be easily used on our algorithms to foil eavesdropping attacks on the wireless channel between the users and the IS, in this section, we focus on the inference attack performed by the active adversary.

In a mobile social network, users typically interact with

Algorithm 1: Evaluate location privacy

Input: User's privacy requirement
$PR(k) =< (r_1, l_1, s_1), \cdots, (r_i, l_i, s_i), \cdots, (r_n, l_n, s_n) >$,
User's previous check-in location l_{i-1} and check-in time t_{i-1}
User's Current location l_i and current time t_i
Output: probable leaked hidden POIs, together with the leakage probability

// Find the candidate POIs the user has visited $\triangle t = t_i - t_{i-1}$;
if $\triangle t \leq dis(l_{i-1}, l_i/v_{max})$ **then**
  $L_{pre} = \emptyset$;
end
else
  $L_{pre} = \{l_m | dis(l_{i-1}, l_m) + dis(l_m, l_i) \leq \triangle t * v_{max}\}$;
end
// Find the candidate POIs that user will visit L_{near} = {the top k nearest POI to l_i};
L_{maj} = {the top k POIs that other users will visit};
L_{rate} = {u_k's top k highest rated POIs };
$L_{next} = L_{near} \cup L_{maj} \cup L_{rate}$;
//Evaluate whether the users' privacy requirement is satisfied
if $(LS(k) \cap (L_{pre} \cup L_{next})) = \emptyset$) **then**
  return NULL;
end
else
  for *each* l_j *in* $(LS(k) \cap L_{pre})$ **do**
    if $p_{nor}(l_j | l_{i-1}, l_i, u_k) > s_j$ **then**
      $H_{pre} = H_{pre} \cup < l_j, p_{nor}(l_j | l_{i-1}, l_i, u_k) >$;
    end
  end
  for *each* l_j *in* $(LS(k) \cap L_{next})$ **do**
    if $p_{nor}(l_j | l_i, u_k) > s_j$ **then**
      $H_{next} = H_{next} \cup < l_j, p_{nor}(l_j | l_i, u_k) >$;
    end
  end
end
return $H_{pre} \cup H_{next}$;

friends via geo-location content. Ideally, the adversary cannot find the user's hidden location when the users disable the location service in locations that they regard as sensitive. However, a strong adversary may know the history data of a particular user. When this data is combined with side information, the adversary can use this information to perform inference attacks. The goal of the adversary is to improve its probability of successfully guessing the real hidden location from the user's current location. In our algorithms, the inference attack is avoided by using obfuscation, which can be achieved by deleting the locations that can be used to infer users' location privacy. We first calculate the possible POIs that the users will visit or have visited, and we then made sure that for each of the candidate POIs, the probability that the user will visit (has visited) the POI is below the privacy requirement for the POI. Thus, although the adversary finds that the user has visited (will visit) other POIs, the adversary cannot use a user's check-in locations to reveal that user's hidden location. As a result, it is meaningless for adversaries to reverse our scheme, which can effectively protect the user's location privacy.

6. Performance Evaluations

In this section, we conduct intensive experiments to evaluate our inference model and privacy-aware location-sharing framework. We present experimental settings in Sect. 5.1 and analyze the experimental results in Sect. 5.2.

Table 1 Statistics of the two datasets

Statistic	Foursquare	Yelp
Number of users	121,03	51,714
Number of POIs	78,425	13,916
Number of check-in or ratings	934,531	272,834
Number of social links	136,334	285,219

6.1 Simulation Setup

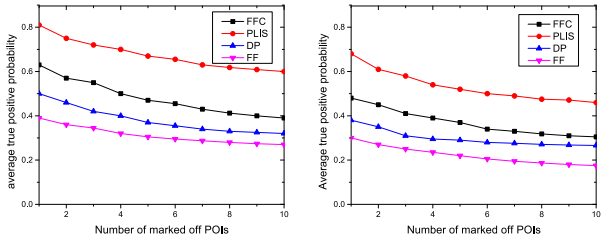
We evaluate our system using the real-world dataset we created using Foursquare [15] and Yelp [21]. We have the check-in or rating history, the corresponding check-in time, and the social ties of each user in the two datasets. Table 1 shows the statistics of the data set. Note that the Foursquare data provide the check-in frequencies of users to POIs, which can be used to calculate the implicit rating, whereas the Yelp data set offers the explicit rating of the users' POIs. Realistically, we can only use the check-in history to predict future check-in behaviors. Thus, we use the half of the check-in data with the earlier check-in time as the training set and the other half of the check-in data as the testing set. In addition, our proposed method PLIS, which is described in Sect. 5, is compared with three existing schemes, DP, FF and FFC, which represent the inference models used in [12], [13] and [14], respectively.

6.2 Experimental Results

6.2.1 Study on Accuracy

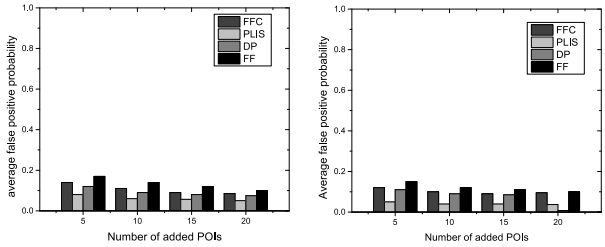
The sensitive locations are invisible in the two datasets. Thus, we artificially generate two sensitive location sets for each user in our experiment. Given a user u_k, hidden location set A is generated by randomly marking off a portion of the POIs where he has checked in; hidden location set B is generated by adding POIs that are geographically located between l_{i-1} and l_i where u_k did not check in. We then use the following metrics to evaluate our system: (i) true positive probability, which refers to the average probability that the locations in hidden location set A are identified as the hidden locations; and (ii) false positive probability, which refers to the average probability that the locations in hidden location set B are identified as the hidden locations.

Figure 6 shows the effects of the number of marked POIs on the average true positive probability, whereas Fig. 7 depicts the effects of added POIs on false positive probability. In Fig. 6, the higher the true positive probability, the better the approach, whereas the false positive probability in Fig. 7 represents the opposite. Our proposed method (PLIS) exhibits the best performance in terms of both true and false positive probabilities. DP only factors in the majority of users' primary historical check-in sequences to predict a



(a) Foursquare (b) Yelp

Fig. 6 Performance evaluations on true positive probability

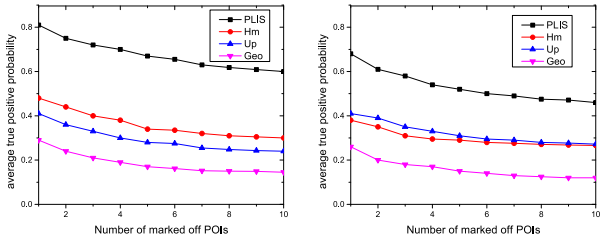


(a) Foursquare (b) Yelp

Fig. 7 Performance evaluations on false positive probability

user's future destinations. With this approach, it is difficult to evaluate whether the user has visited a hidden location if the user's check-in sequences are personal and unpopular. It should be noted at this juncture that FF has both low true positive and false positive probabilities. This result stems from a data sparsity problem when few of a user's friends have visited a hidden location within the given time span. FFC uses the check-in history of a user's friends and also uses common interests to predict the user's hidden locations. It generates the second best prediction results. However, FFC still ignores the geographical distribution in the users' check-in history. In addition, it is inadvisable to assign the same weight to all users' interests (preferences). Some users are affected more by common interests, whereas others may rely more on the geographical influence or on the patterns of the majority of users. Another drawback of FFC is that it has a high false positive probability, particularly in dense urban areas with few check-ins.

Our proposed method exhibits the best performance in terms of both true positive probability and false positive probability. In particular, PLIS realizes significant improvement in comparison with the other prediction techniques. The reason is threefold: (1) we model a personalized geographical check-in distribution for each user based on the user's check-in history; (2) we take full advantage of the user's preferences by seamlessly combining the bias of users and the popularity of POIs into a reasonable relevance score based on the TF-IDF scheme; and (3) we combine geographic information, human mobility patterns, and user preferences to model a personalized prediction for each user.



(a) Foursquare (b) Yelp

Fig. 8 Performance evaluations on different factors

6.2.2 Study of the Three Factors of Our Method

In this section, we study the three factors in PLIS: geographic information, human mobility patterns, and user preferences, referred to as Geo, Hm, and Up, respectively. Figure 8 plots the prediction accuracy of the three factors based on Eqs. (2), (4), and (11), respectively. We have two observations: (i) all three factors play an important role in PLIS for hidden location prediction and compete with one another. For example, Hm outperforms Up on the Foursquare data, but the reverse is true regarding the Yelp data; (ii) integrating the three components is helpful in increasing the prediction accuracy. As shown in Fig. 8, PLIS is significantly superior to each factor alone because geographic information, human mobility patterns, and user preferences affect people to different degrees in the real world. Thus, the hidden location prediction model makes the best use of various types of valuable information implied by users' check-in behaviors at POIs.

6.2.3 Privacy-Utility Tradeoff

In this section, we study the price, in terms of utility, that users must pay for a formal privacy guarantee. We know that users in the mobile social network always want to publish as much geo-location content as required to receive better service. Therefore, we propose a performance metric, called the utility ratio, to evaluate the users' utility. It is calculated as follows:

$$Utility = \frac{|L_{can}|}{|L_{want}|} \tag{19}$$

where L_{want} denotes the set of the POIs at which the user wants to check-in, and L_{can} denotes the POIs where the user can check-in without damaging his/her location privacy.

We randomly select a portion of the POIs (5 percent of the POIs and 10 percent of the POIs) in the dataset and mark them as sensitive locations at which no users will check-in. For purposes of simplicity, in this experiment, we assume that the privacy requirement for the selected sensitive POIs is the same for all users. However, PLIS allows users to define personal privacy requirements for different POIs.

We vary users' privacy requirements for POIs by changing the value of confidence s_i, which is defined above in Sect. 3.3 and denotes that no adversary can deduce that

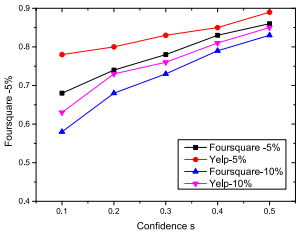


Fig. 9 Performance on utility ratio

the probability of the user visiting l_i is higher than s_i. As shown in Fig. 9, the performance degradation of our system utility is small. For example, when the ratio of the sensitive POIs is 0.10, and the required confidence level s_i is 0.3, the utility ratio of PLIS in Foursquare data is 0.73. This is because the false positive probability of our approach is low, while the true positive probability of our approach is high. Thus, PLIS only suppresses the POIs that are relevant to the sensitive POIs. The result implies that PLIS can offer strong privacy guarantees (by choosing a smaller value of confidence s) without sacrificing too much utility.

7. Conclusion

Privacy-aware location information management is an important problem in LBSs. In this paper, we proposed a new location privacy attack called the hidden location inference attack. By means of data analysis, we discovered three factors that can be used to infer a user's private location: geographic information, human mobility patterns and user preferences. Then, we evaluated the separate impact of each of these factors on users' check-in behaviors. Moreover, we proposed a novel algorithm that is based on Multiple Linear Regression to integrate the above heterogeneous factors and provide a personalized check-in behavior prediction for each user. Based on the predictor model, we proposed a privacy-aware location-sharing framework that warns users when their actual sharing behaviors do not match their sharing rules. Finally, we evaluated our scheme using two real-world datasets. The results of our experiment demonstrate the validity and practicality of the proposed strategy.

Based on the work presented here, future research will address the following issues. First, the adversary may use different models to infer the users' location privacy, this may degrade the performance of PLIS mechanism. An extension mechanism to resist these attack models will be proposed. Second, because PLIS mechanism does not assume all side information, there are also plans to study more types of *side information*(e.g.,social relationship,gender) to enhance the effectiveness of the proposed mechanism.

Acknowledgments

This research was partially supported by the National Basic Research Program of China under grant No. 2011CB302605, No. 2013CB329602, and the National Science Foundation of China (NSF) under grants No. 61202457, No. 61402137, and No. 61402149.

References

[1] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," Proc. 1st international conference on Mobile systems, applications and services, pp.31–42, ACM, 2003.
[2] H.P. Li, H. Hu, and J. Xu, "Nearby friend alert: Location anonymity in mobile geosocial networks," Pervasive Comput., IEEE, vol.12, no.4, pp.62–70, 2013.
[3] Y. Wang, D. Xu, X. He, C. Zhang, F. Li, and D. Xu, "L2p2: Location-aware location privacy protection for location-based services," INFOCOM, 2012 Proceedings IEEE, pp.1996–2004, IEEE, 2012.
[4] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Achieving k-anonymity in privacy-aware location-based services," INFOCOM, 2014 Proceedings IEEE, pp.754–762, IEEE, 2014.
[5] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Enhancing privacy through caching in location-based services," Proc. of IEEE INFOCOM, 2015.
[6] B. Palanisamy and L. Liu, "Mobimix: Protecting location privacy with mix-zones over road networks," Data Engineering (ICDE), 2011 IEEE 27th International Conference on, pp.494–505, IEEE, 2011.
[7] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," INFOCOM, 2012 Proceedings IEEE, pp.972–980, IEEE, 2012.
[8] X. Zhao, L. Li, and G. Xue, "Checking in without worries: Location privacy in location based social networks," INFOCOM, 2013 Proceedings IEEE, pp.3003–3011, IEEE, 2013.
[9] X.Y. Li and T. Jung, "Search me if you can: privacy-preserving location query service," INFOCOM, 2013 Proceedings IEEE, pp.2760–2768, IEEE, 2013.
[10] M.E. Andrés, N.E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," Proc. 2013 ACM SIGSAC conference on Computer & communications security, pp.901–914, ACM, 2013.
[11] T. Murakami, A. Kanemura, and H. Hino, "Group sparsity tensor factorization for de-anonymization of mobility traces," Trustcom/BigDataSE/ISPA, 2015 IEEE, pp.621–629, IEEE, 2015.
[12] A.Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pp.254–265, IEEE, 2013.
[13] A. Sadilek, H. Kautz, and J.P. Bigham, "Finding your friends and following them to where you are," Proc. fifth ACM international conference on Web search and data mining, pp.723–732, ACM, 2012.
[14] Z. Huo, X. Meng, and R. Zhang, "Feel free to check-in: Privacy alert against hidden location inference attacks in geosns," Database Systems for Advanced Applications, pp.377–391, Springer, 2013.
[15] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks.," ICWSM, 2012.
[16] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," ICwSM, vol.11, pp.70–573, 2011.
[17] J. Ramos, "Using tf-idf to determine word relevance in document queries," Proc. first instructional conference on machine learning, 2003.
[18] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," Internet Comput., IEEE, vol.7, no.1, pp.76–80, 2003.
[19] L.S. Aiken, S.G. West, and S.C. Pitts, "Multiple linear regression," Handbook of Psychology, Wiley, 2003.
[20] W.W. Chin, "The partial least squares approach to structural equation modeling," Modern Methods for Business Research, vol.295,

no.2, pp.295–336, 1998.
[21] Yelp. http://www.yelp.com/dataset_challenge. Accessed June 19, 2015.

Zhikai Xu received the B.S. degree in Computer science from Dalian University of Technology (DUT), Dalian, China in 2009, and his M.S. degree in Computer Science form Harbin Institute of Technology, Harbin, China in 2011. From 2011 to now, he has been work as a Ph.D. candidate in Department of Computer Science and Technology at Harbin Institute of Technology, Harbin, China. His research interests include Mobile Cloud computing and security.

Hongli Zhang received her B.S. degree in Computer Science from Sichuan University, Chengdu, China in 1994, and her Ph.D. degree in Computer Science from Harbin Institute of Technology (HIT), Harbin, China in 1999. She is a professor in Computer Science and Technology Department of HIT and the vice director of National Computer Information Content Security Key Laboratory. Her research interests include network and information security, network measurement and modeling, and parallel processing.

Xiangzhan Yu was born in 1973. He is currently a professor in School of Computer Science and Technology in HIT. His research interests include Peer-to-Peer system, cloud computing and information security.

Shen Su received the B.S. and M.S. degree in Computer Science from Harbin Institute of Technology, Harbin, China, from 2004 to 2010. From 2010 to now, he has been work as a Ph.D. candidate in Department of Computer Science and Technology at Harbin Institute of Technology (HIT), Harbin, China. His research interests include communication and mobile computing.