JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 803-820 (2018)
DOI: 10.6688/JISE.201807_34(4).0001

PTPP: Preference-Aware Trajectory Privacy-Preserving over Location-Based Social Networks

LIANG ZHU[1,3,+], HAIYONG XIE[2], YIFENG LIU[2], JIANFENG GUAN[3],
YANG LIU[3] AND YONGPING XIONG[3]
[1]*School of Computer and Communication Engineering*
Zhengzhou University of Light Industry
Zhengzhou, 450001 P.R. China
[2]*Innovation Center*
China Academy of Electronics and Information Technology
Beijing, 100041 P.R. China
[3]*State Key Laboratory of Networking and Switching Technology*
Beijing University of Posts and Telecommunications
Beijing, 100876 P.R. China
E-mail: {lzhu; jfguan; liu.yang; ypxiong}@bupt.edu.cn; haiyong.xie@ieee.org; yliu@csdslab.net

Trajectory privacy-preserving for Location-based Social Networks (LBSNs) has received much attention to protect the sensitive location information of subscribers from leaking. Existing trajectory privacy-preserving schemes in literature are confronted with three problems: (1) it is limited for privacy-preserving by only considering the location anonymization in practical environment, and the sensitive locations are always revealed by this way; (2) they fail to consider the user preference and background information in trajectory anonymization, which is important to keep personalized location-based service; (3) they cannot be adapted to different kinds of privacy risk levels, resulting in low the service precision. To tackle the above problems, we propose PTPP, a preference-aware trajectory privacy-preserving scheme. First, we model the user preference by considering geographical information, semantical information, movement pattern, user familiarity and location popularity. Then, we classify the privacy risk levels according to user familiarity and location popularity. Finally, we propose a preference-aware trajectory anonymization algorithm by considering privacy risk levels. The experimental results show that our method outperforms a state-of-the-art trajectory privacy-preserving method in terms of data utility and efficiency.

Keywords: location-based social networks, trajectory privacy-preserving, movement pattern, user preference, behavior analysis

1. INTRODUCTION

Recently, Location-Based Social Networks (LBSNs) have received great attention due to the rapid development of online social networks and physical localization technologies. In LBSNs, people make use of sensor-embedded mobile devices to share the location-related contents with their social friends, and publish the geographical locations to LBSN servers in order to obtain different kinds of location-based services [1, 2]. For example, *Loopt* provides the service for smart-phone subscribers to share their locations selectively with other people [3]; *Twitter* has the ability to connect the corresponding location information while subscribers share interested contents with their friends [4];

Received June 14, 2017; accepted July 22, 2017.
Communicated by Gabriel-Miro Muntean.
[+] Corresponding author.

Foursquare and *Gowalla* can recommend the personalized service to subscribers by analyzing the collected "check-in" information of point-of-interests (POIs). Therefore, subscribers would know the latest news about some places at any time, and find the nearest friends around them [5].

However, privacy leakage will inevitably happen when subscribers enjoy the great convenience provided by LBSNs. Untrusted third party may steal users' data information to do illegal activities. For subscribers, the complete Global Positioning System (GPS) trajectory may not be published because of privacy consideration. But it can be easily deduced by attackers according to the time-space relativity of geographical locations. By analyzing the GPS trajectory, attackers can acquire the personalized privacy information of subscribers, such as family address, working place and living habit, *etc.* Even the next location can be predicted by deduced movement pattern, which seriously affects the security of subscribers.

In this paper, we study the problem of trajectory privacy-preserving when one subscriber publishes his/her original trajectory information to LBSN servers. The key challenges of our proposal are *how to efficiently extract movement pattern from subscribers' trajectories*, *how to classify the privacy risk levels according to different preference on locations and how to make a trajectory anonymization scheme to satisfy the personalized demands of subscribers*?

Trajectory privacy-preserving is a new type of privacy protection task that comes along with LBSNs. Different with traditional location privacy-preserving methods, such as false location, *k*-anonymity or encryption, it concerns more about protecting the leakage of sensitive locations which reflect the interest and preference of subscribers. There are mainly three methods to protect the trajectory privacy: fake data [6], spatial cloaking [7] and inhibition technique [8]. Fake data method is to add some false location information when subscribers publish the original GPS trajectory to servers. Spatial cloaking can generalize the sensitive location of original GPS trajectory to reduce the possibility that attackers find the accurate location information. Inhibition technique aims to forbid the data release of sensitive location information to protect the individual privacy of subscribers.

Although the growing interest in privacy-preserving has resulted in thousands of peer-reviewed publications, there is still significant ongoing work addressing many challenges. There are three problems for current privacy-preserving schemes. First, it is limited for privacy-preserving by only considering the location anonymization in practical environment, and the sensitive locations are always revealed by this way. Second, they fail to consider the user preference and background information in trajectory anonymization, which is important to keep personalized location-based service. Third, they cannot be adapted to different kinds of privacy risk level, resulting in low service precision.

Therefore, we propose a Preference-aware Trajectory Privacy-Preserving (PTPP) scheme, which can not only protect sensitive locations from leaking, but also provide effective service for users. This study is a significant extension for our previous work [9] by constructing the detailed attack model and proposing an extended user preference-aware method. The main contributions of this paper can be summarized as follows:

(1) We utilize two clustering process to obfuscate the original GPS position, and model

the user preference by considering geographical information, semantical information, movement pattern, user familiarity and location popularity.

(2) By classifying the privacy risk level according to user familiarity and location popularity, we propose a preference-aware trajectory anonymization algorithm.

(3) We conduct an experiment to evaluate the validity and efficiency of the proposed PTPP.

2. RELATED WORK

In this section, we briefly review some related works on trajectory privacy-preserving. The main objective of this paper is to present a preference-aware trajectory privacy-preserving scheme (PTPP), in order to adaptively protect user privacy from leaking.

(A) Fake trajectory

Fake location is always used for location privacy-preserving. By making use of fake location, users can obtain the satisfied service while their actual locations will not be published [10]. Likewise, the method of fake trajectory can also be used to protect user trajectory privacy from leaking. Utilizing fake trajectory, the real trajectory can be concealed, which decreases the risk of privacy disclosure. Yiu *et al.* [11] proposed a SpaceTwist framework to make a balance among location privacy, query performance and query accuracy. In order to save the energy consumption of resource-constrained mobile devices, Liu *et al.* [12] proposed a strategy selection algorithm modeled by Bayesian games. For trajectory privacy-preserving, You *et al.* [13] proposed a method of random and rotation pattern to generate dummy trajectories. After that, a dummy-based anonymization scheme according to different user movement trajectory was proposed by Kato *et al.* [14]. However, due to more noises in the query of fake trajectory, more redundant results may still be returned. Therefore, it made a higher communication cost in mobile client.

(B) Spatial cloaking

Spatial cloaking can well reduce the load of mobile client by making use of a Trust Third Party (TTP). According to spatial cloaking, k-anonymity is the popular method for trajectory privacy-preserving. The first k-anonymity model proposed by Sweeney [15] was applied to the privacy-preserving of relational data base. After that, Gedik *et al.* [16] proposed a unified privacy personalization framework to guarantee the location privacy in TTP. In order to provide a formal guarantee for the strength of k-anonymity, Kalnis *et al.* [17] proposed a scheme to model the suitable anonymizing regions and discussed the trade-offs. For trajectory k-anonymity, Gao *et al.* [18] proposed a framework, named TrPF, to obfuscate the original trajectory by making use of $k-1$ similar trajectories. Also, Han *et al.* [19] proposed a semantic space translation algorithm (SST) to balance the data utility and k-anonymity trajectory privacy-preserving in TTP. The method of trajectory partitioning was proposed by Shin *et al.* [20] for trajectory anonymity. However, spatial cloaking relies on the precondition that TTP is fully trusted. In practice, the TTP is still vulnerable if attackers urgently want to acquire the real trajectory information of victims.

(C) Inhibition technique

Inhibition technique is to restrain some data of original trajectory not to be published, which can effectively protect the sensitive location from leaking. The key problem for inhibition technique is that how to find the sensitive location to be suppressed while ensure the data utility. Terrovitis *et al.* [21] proposed an inhibition-based trajectory privacy-preserving method according to the partial trajectory obtained by attackers. Another inhibition-based trajectory privacy-preserving scheme was proposed by Gruteser *et al.* [22]. This scheme divides the geographical region into sensitive region and non-sensitive region. Once the object moved into sensitive region, inhibition technique was activated to protect trajectory privacy. Also, Chen *et al.* [23] utilized location inhibition technique to construct a tailored privacy model for trajectory privacy-preserving. Inhibition technique is an efficient method to protect trajectory privacy in the condition that the special background information of attackers is known. However, the background information of attackers is hard to be acquired in practice. In this paper, attackers are supposed that they can infer the preference information of victims according to the uploaded trajectory data. Therefore, we proposed a preference-aware trajectory privacy-preserving scheme (PTPP) to protect user privacy according to the difference preference for locations.

3. OVERVIEW OF PTPP ARCHITECTURE

In PTPP, the original trajectory will be processed by a Trust Third Party (TTP) before publishing. Fig. 1 shows the architecture and work flow of PTPP, which involves four phases for trajectory anonymization: (1) stay-point extraction; (2) location extraction; (3) movement pattern extraction and (4) privacy risk level retrieval. The privacy-preserving scheme runs in the phase of trajectory information publishing. For privacy-preserving, we construct the user preference model and then classify the privacy risk level according to user preference and background information. Finally, the adaptive trajectory anonymization algorithm is performed according to different privacy risk level. In a nutshell, the operation of trajectory anonymization includes three steps. The detailed workflow of PTPP is summarized as follows.

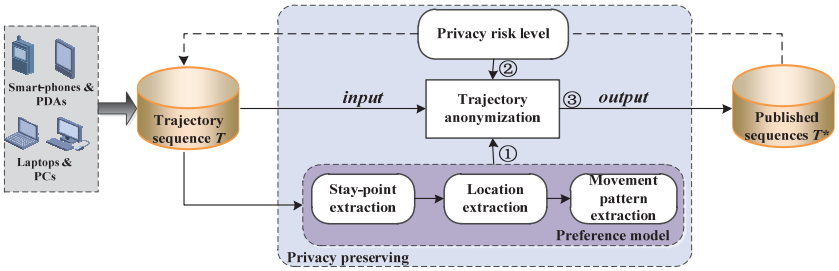


Fig. 1. The architecture and work flow of PTPP.

Step 1: The original positions are firstly clustered into stay-points by stay-point extraction then the stay-points are clustered into locations by location extraction. After stay-point extraction and location extraction, the original positions can be transformed into

different locations, in order to protect the actual positions of users from leaking. What's more, the movement pattern is extracted to reflect the interest or preference of users. Through stay-point extraction, location extraction and movement pattern extraction, the time-varying preference model of each user is constructed.

Step 2: Although stay-point extraction and location extraction can protect the accurate position information from leaking, the movement pattern may be inferred by attackers, which leads to the interest or preference of users may be revealed. Therefore, different privacy risk levels are classified according to the generated preference model.

Step 3: There are three methods with different privacy-preserving degree to protect the sensitive location information from leaking. For one trajectory, it is better to utilize adaptive privacy-preserving methods instead of single privacy-preserving methods. Therefore, trajectory anonymization is necessary to output the obfuscated trajectory sequence.

To further explain these main features in PTPP, we detailedly describe the scheme designs and algorithms in the next two sections.

4. PTPP SCHEME DESIGNS

In this section, we first give some description for the problem definition. Then, we explain the adversary model in PTPP. After that, the user preference model is constructed. Finally, the privacy risk level is given according to the preference model of each user. Table 1 lists the relevant notations and definitions.

Table 1. Notations and definitions.

Notations	Descriptions
p	The raw position
s	The stay-point
$\mathcal{S}$	The set of stay-point
L	The location
Tra_L	The location trajectory
Tra_C	The type trajectory
$\mathbb{M}^C$	The user-location matrix
$F_{u_i}^C(n)$	The user familiarity
$P_{L_i}^C(n)$	The location popularity
T	The raw trajectory sequence
T^*	The anonymous trajectory sequence
$\mathbb{Z}$	Anonymous zone
$\mathbb{Z}_i^S$	The anonymous zone of stay-points
$\mathbb{Z}_j^L$	The anonymous zone of locations

4.1 Problem Definition

Trajectories of mobile users are constructed by some successive locations in a cer-

tain time interval. Trajectory privacy-preserving is to protect the sensitive locations or personalized privacy information from leaking. In this paper, we take the mobile GPS position of each user as point. And the original trajectory can be defined as: $Tra\_p = p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$, where p_i is the real position visited by user.

Definition 1: Position p: is a two-tuple as $\langle lon, lat \rangle$, which represents the longitude and latitude of the position.

Each GPS position records the real footprint of people. Users can publish their historical positions to LBSNs server to enjoy the personalized service.

Definition 2: Stay-point S: is a three-tuple as $\langle lon, lat, \theta_t \rangle$, where $\langle lon, lat \rangle$ is the coordinate of stay-point, θ_t is the residence time.

Stay-point is the cluster of positions. It represents a geographical region where users stay for a while, and describes that the users have performed a meaningful activity. For example, users enter into a building where the satellite signal is weak, such as shopping center, movie theater, museum, *etc.* Another example is that users wander in an external geographical region, and not just go through this region, *e.g.* tourist attractions.

Definition 3: Location L: is a three-tuple as $\langle lon, lat, type \rangle$, where $\langle lon, lat \rangle$ is the coordinate of location, `type` is the corresponding semantic information.

Location represents a geographical region clustered by the stay-points with the same semantic information. It not only includes these stay-points on semantic space, such as school, shopping mall, restaurant, *etc.*, but also has nothing with time.

Definition 4: Anonymous zone $\mathbb{Z}$: is a two-tuple as $\langle k, l \rangle$, where k is the number of locations included by the anonymous zone, l is the privacy-preserving level.

Anonymous zone reflects the degree of privacy-preserving through adjusting the size of zone. Also it can measure data utility to explain how much information has been lost during the stage of trajectory anonymization.

4.2 Adversary Model

In this paper, we make an assumption that LBSNs server always provides the "public and honest" services to subscribers. In LBSNs, it has the ability to infer subscribers' sensitive location information by analyzing the published location-related contents. However, attackers may easily acquire the sensitive information of subscribers through LBSNs server. Even attackers can mine the movement pattern and preference of subscribers according to the obtained sensitive information.

In order to construct the adversary model, the attackers are assumed to know some *priori* knowledge about the semantic information of a victim's actual location L_i. In addition, the trajectory with time series can be acquired.

According to the semantic information of each location, semantic trajectory sequence can be generated as $C_1 \rightarrow C_2 \rightarrow \ldots \rightarrow C_n$. The preference or interest of victims can be acquired through mining the frequent subsequences of semantic trajectory sequence. For example, if attacker discovers that victim A owns the frequent subsequences as *School* $\rightarrow$ *Sport* $\rightarrow$ *Restaurant*. When victim A has visited the location sequence as *School* $\rightarrow$ *Sport*, the attacker may infer that victim A will visit the location of "*Restau-*

rant" with a high probability.

Therefore, it exists security threat for users when they experience the personalized service provided by LBSNs server. Our PTPP scheme processes the original trajectory sequences of subscribers before being published to LBSNs server. Based on the trajectory anonymization algorithm, we can protect the sensitive location information from leaking.

4.3 Preference Model

(A) Stay-point and location extraction

We utilize two clustering processes to hide the original data information. It can be divided into two steps, *i.e.*, stay-point extraction and location extraction.

For stay-point extraction, a stay-point S_i can be computed by Eq. (1).

$$S_i(lon) = \sum_{i=m}^{n} \frac{p_i(lon)}{\left|\mathbb{Z}_i^S\right|}, S_i(lat) = \sum_{i=m}^{n} \frac{p_i(lat)}{\left|\mathbb{Z}_i^S\right|}, \tag{1}$$

where $\mathbb{Z}_i^S$ stands for the anonymous zone of stay-points, $p_j(lon)$ and $p_j(lat)$ represent the longitude and latitude of each raw point p_j, respectively. And we can generate the obfuscated trajectory consisted by the stay-points as $Tra\_S = S_1 \to S_2 \to \ldots \to S_n$.

For location extraction, the longitude and latitude of location L_j can be computed by Eq. (2).

$$L_j(lon) = \frac{\sum_{S_i \in \mathcal{S}_j} S_i(lon)}{\left|\mathbb{Z}_j^L\right|}, L_j(lat) = \frac{\sum_{S_i \in \mathcal{S}_j} S_i(lat)}{\left|\mathbb{Z}_j^L\right|}, \tag{2}$$

where $\mathbb{Z}_j^S$ stands for the anonymous zone of locations, $S_i(lon)$ and $S_i(lat)$ represent the longitude and latitude of each stay-point S_i, respectively. And we can generate the obfuscated trajectory consisted by the locations as $Tra\_L = L_1 \to L_2 \to \ldots \to L_n$.

The detailed clustering algorithm can be referred to our previous studies [24, 25]. In this paper, we mainly study the trajectory anonymization based on the generated location trajectory. According to the above two clustering processes, attackers cannot easily acquire the accurate positioning information of users. However, the frequent movement patterns can be mined from the successive location sequence, which inevitably reveals the living habits and daily routines of users. Therefore, the movement patterns must be taken into consideration to protect trajectory privacy.

(B) Geographical information

In geographical space, the distance between two locations can well reflect users' visitation behaviors. Generally speaking, the larger the distance is, the smaller the probability that one user visits the next location is. The correlation between two locations decreases as the distance between them increases. Therefore, by making use of Gauss formula, the distance similarity between two locations can be computed as:

$$Sim_{geo}(L_i, L_j) = \exp\left(-\frac{D(L_i, L_j)^2}{2}\right), \tag{3}$$

where $D(L_i, L_j)$ is the Euclidean distance between L_i and L_j.

Also, the habit and movement pattern of users can be mined from their historical trajectories, which has been demonstrated by our previous studies. Thus, the next location that user u_k will visit corresponding to current location can be predicted through the historical trajectory. Let $His(u_k) = \{L_1, L_2, \ldots, L_n\}$ denote the historical trajectory of user u_k, L_i is the current location. The probability that user u_k will visit L_j after L_i can be computed as:

$$P_{geo}(L_j \mid L_i, u_k) = aP_{geo}(L_i, L_j) + (1-a) \sum_{L_k \in His(u_k)} P_{geo}(L_k, L_j), \tag{4}$$

where a is an activation parameter ranging within [0,1].

(C) Semantical information

We define the type of each location by making use of Term Frequency-Inverse Document Frequency (TF-IDF). The weight of each type i for a stay region is computed as:

$$w_k = \frac{n_k}{N} \times \log \frac{\left|\mathbb{Z}^S\right|}{\left|\mathbb{Z}_k^S\right|}, \tag{5}$$

where N is the total number of positions appear in the region, n_i is the number of positions for type i, and $|\mathbb{Z}^S|$ is the number of anonymous zone of stay-points, $|\mathbb{Z}_k^S|$ is the number of anonymous zone of stay-points with type k. And the feature vector of each stay-point can be defined as $\mathbf{f}_s = \langle w_1, w_2, \ldots, w_n \rangle$. We make use of the number of non-zero weight of each type to compute the feature vector of each location. The weight of each type i for a location is computed as:

$$\overline{w}_i = \frac{\sum_{w_i \in \mathbf{f}_s} w_i}{\left|\{w_i \mid w_i > 0\}\right|}. \tag{6}$$

And the feature vector of each location can be defined as $\mathbf{f}_L = \langle W_1, W_2, \ldots, W_n \rangle$.

In this paper, we select the type with the highest weight value as the semantic description of each location. Therefore, the semantic trajectory corresponding to location trajectory can be represented as $Tra\_C = C_1 \rightarrow C_2 \rightarrow \ldots \rightarrow C_n$.

(D) Movement pattern extraction

In order to extract the movement pattern, we let n denote the length of pattern, and ρ denote the occurrence number. **Algorithm 1** shows the detailed process of movement pattern extraction.

Algorithm 1: Movement Pattern Extraction

Input: semantic trajectory *Tra_C*, length of pattern n, occurrence number ρ

Output: set of movement pattern P //The last bit of each pattern is occurrence number

1. **Initialize** $P = \varnothing$, $num = 0$, $i = 1$;
2. **Define** l is the length of semantic trajectory *Tra_C*;
3. **While** $i <= (l-n+1)$

```
4.      extract the n-length sequence seq begin with i;
5.      If seq belongs to Tra_C // the sequence is occurred in one day
6.        If seq doesn't belong to P
7.              num=1;
8.              put seq into P;
9.        Else
10.             add num of sequence that same with seq in P;
11.       End if
12.     End if
13.     i = i + 1;
14. End while
15. Delete the element of P if num < ρ;
16. Sort the elements of P according to num;
```

(E) User familiarity and location popularity

In order to well reflect the user preference and background information for trajectory privacy-preserving, we formulate the user familiarity for each type and the location popularity for each location in the corresponding type. Based on the thought of Hypertext Induced Topic Search (HITS) [26], users and locations are considered as hub nodes and authority nodes, respectively. The number of hub nodes is user familiarity, and the number of authority nodes is location popularity. Thus, user familiarity can be computed by the sum of the value of authority nodes and location popularity can be computed by the sum of the value of hub nodes.

According to the different semantic information, we classify the location in order to get the user-location matrix $\mathbb{M}^L$, and each element M_{ij}^L represents the visit frequency of location j for user i. Also, the user familiarity with each type C is defined as $F_{u_i}^C(n)$, and the location popularity with these type is defined as $P_{L_j}^C(n)$. Corresponding to each type, the user familiarity can be computed as:

$$F_{u_i}^C(n) = \sum_{L \in C} (M_{ij}^L \times p_{L_j}^C(n-1)), \tag{7}$$

and the location popularity can be computed as:

$$P_{L_j}^C(n) = \sum_{u} (f_{u_i}^C(n-1) \times M_{ij}^L). \tag{8}$$

In this paper, we define $\mathbf{F}_n^C$ as the vector of user familiarity with type C, and $\mathbf{P}_n^C$ as the vector of location popularity with the same type C. By making use of iterative method, $\mathbf{F}_n^C$ and $\mathbf{P}_n^C$ can be computed as:

$$\mathbf{F}_n^C = \mathbb{M}^L \bullet (\mathbb{M}^L)^T \bullet \mathbf{P}_{n\text{-}1}^C \tag{9}$$

$$\mathbf{P}_n^C = (\mathbb{M}^L)^T \bullet \mathbb{M}^L \bullet \mathbf{F}_{n\text{-}1}^C, \tag{10}$$

where n represents the iterative number.

Initially, $\mathbf{F}_0^C = \mathbf{P}_0^C = (1, 1, \ldots, 1)^T$, and the process stops until

$$\|\mathbf{F}_n^C - \mathbf{F}_{n\text{-}1}^C\| + \|\mathbf{P}_n^C - \mathbf{P}_{n\text{-}1}^C\| < \epsilon. \tag{11}$$

In this way, we have modeled the user preference by considering movement pattern, user familiarity and location popularity.

4.4 Privacy Risk Level

Traditional trajectory anonymization always makes a uniform scheme for all the locations belong to one trajectory, which not only fails to consider the user preference for different locations, but also influences the efficiency of privacy preserving. In this paper, we divide the locations into four privacy risk levels. And different location anonymization methods are adopted to protect the trajectory privacy. We define λ and τ as the threshold of user familiarity and location popularity, respectively. Four privacy risk levels are classified as follows.

(1) *Non-Familiar and Popular* (*NFP*): A location belongs to this region in the condition that user familiar is less than λ, and the location popularity is not less than τ. It means the user is not an expert in the type of the location. The attackers cannot easily deduce the user preference if this location is leaked. Also, the location has been visited by many people because the location popular is high. So we don't need to protect this location information.
(2) *Non-Familiar and Non-Popular* (*NFNP*): A location belongs to this region in the condition that user familiar is less than λ, and the location popularity is less than τ. It means the user is not familiar with the location, but the location popularity is low. The attackers can easily deduce the identity information of user who visited this location. So fake data method is used to protect this location information.
(3) *Familiar and Popular* (*FP*): A location belongs to this region in the condition that user familiar is not less than λ, and the location popularity is not less than τ. It means the user is an expert in the type of this location. The user preference would be revealed if the attackers know the user has visited the location. However, the location has high popularity value. So we can utilize spatial cloaking method to find k locations belong to the same type to protect this location information.
(4) *Familiar and Non-Popular* (*FNP*): A location belongs to this region in the condition that user familiar is not less than λ, and the location popularity is less than τ. It means the attackers can accurately deduce the preference information and identity information of user because of the high user familiarity and low location popularity. So inhibition method is necessary to prohibit the location information release to protect user privacy.

Table 2 shows different methods corresponding to different levels.

Table 2. Classified trajectory privacy-preserving method.

	Privacy-preserving level	Method
1	NFP	N/A
2	NFNP	Fake data
3	FP	Spatial cloaking
4	FNP	Inhibition

5. PTPP ALGORITHMS

This section explicitly details the processing of the proposed PTPP scheme and discusses the trade-off between privacy and data utility.

5.1 Preference-Aware Trajectory Anonymization

Based on the classified privacy risk levels, we propose an adaptive trajectory privacy-preserving scheme to satisfy the personalized demand of users. The preference-aware trajectory anonymization algorithm (**Algorithm 2**) is summarized as below. Five main inputs of our algorithm are: (a) movement pattern p; (b) user familiarity threshold λ; (c) location popularity threshold τ; (d) original trajectory sequence T; (e) anonymous zone $\mathbb{Z} = \langle k, l \rangle$. Basically, we define the length of the trajectory as the number of locations it includes. First, for the first location of trajectory T, we determine the type C it belongs, and compute the value of user familiarity F_u^C and location popularity $P_{L_i}^C$. Next, we make use of four solutions with four conditions to protect the location privacy. Here, the candidate locations are randomly selected by fake data method and spatial cloaking method. In future studies, we will refine the above methods by considering distance, velocity, direction, *etc.* Finally, the anonymous trajectory sequences can be constructed by connecting the candidate locations in a chronological order.

Algorithm 2: Preference-aware Trajectory Anonymization

Input: set of movement pattern P, user familiarity threshold λ, location popularity threshold τ, original trajectory sequence T, anonymous zone $\mathbb{Z}$

Output: set of anonymous trajectory sequences T^*

1. **Initialize** $T^* = \varnothing$, $i = 1, j = 1$;
2. **Define** *len* is the length of original trajectory *sequence*;
3. **While** $i <$ *len*
4. Determine the type C for L_i;
5. Compute the value of F_u^C and $P_{L_i}^C$;
6. **While** $j <= k$
7. **If** $(F_u^C < \lambda)$ && $(P_{L_i}^C >= \tau)$
8. put L_i into T_j^*;
9. **Else if** $(F_u^C < \lambda)$ && $(P_{L_i}^C < \tau)$
10. Randomly select one location belongs to the same type;
11. Put the selected location into T_j^*;
12. **Else if** $(F_u^C >= \lambda)$ && $(P_{L_i}^C >= \tau)$
13. Randomly select one location belongs to the other type;
14. Put the selected location into T_j^*;
15. **Else if** $(F_u^C >= \lambda)$ && $(P_{L_i}^C < \tau)$
16. Delete L_i in T_j^*;
17. **End if**
18. $j = j + 1$;
19. **End while**
20. $i = i + 1$;
21. **End while**

Fig. 2 is an example to explain the idea of preference-aware trajectory anonymization algorithm. The top one shows the original trajectory sequence of one user, and the bottom one shows the anonymous trajectory sequences. Thereinto, location L_1 and L_5 belong to category NFP, location L_2 belongs to category NFNP, location L_3 belongs to category FP, and location L_4 belongs to category FNP.

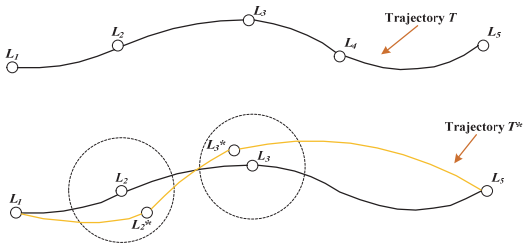


Fig. 2. An example of preference-aware trajectory anonymization algorithm.

5.2 Privacy and Utility Analysis

As we know, the size of the anonymous set is used to tradeoff the degree of privacy for location privacy-preserving. That is to say, the larger the anonymous set, the higher the degree of privacy-preserving. However, for trajectory privacy-preserving, it is not good to quantify the degree of privacy. In this paper, we utilize the theory of Shannon entropy to quantify the trajectory privacy-preserving.

Refer to the definition of information entropy [27], for a group of probability distribution $p_1, p_2, ..., p_n$, the information entropy can be represented as $H = -\sum p_i \log_2 p_i$. We assume the visited location is sensitive at time $t+1$, it has $k-1$ candidate locations at time t. The possibility that user visited one of the k locations at time $t+1$ is defined as $p_1, p_2, ..., p_k$. And we make the possibility that user stayed where he/she was at time t as p_0. So the entropy of trajectory privacy-preserving at time $(t, t+1)$ can be computed as:

$$H_{(t,t+1)} = -\sum_{i=1}^{k} p_i \log p_i - p_0 \log p_0 . \tag{12}$$

According to the feature of entropy, the maximum value is obtained when the possibility that user visited all the candidate locations at time $t+1$ is the same. So we can obtain

$$MaxH_{(t,t+1)} = -\log\left(\frac{1}{k+1}\right). \tag{13}$$

where it includes the location that user still stayed while he/she was at time t.

Thus, we can tradeoff the degree of trajectory privacy-preserving as:

$$H_{\%} = \frac{H_{(t,t+1)}}{MaxH_{(t,t+1)}} 100\%. \tag{14}$$

The larger the value of $H_{\%}$ is, the higher the degree of trajectory privacy-preserving.

6. PERFORMANCE EVALUATION

In this section, we conduct an experiment and evaluate the performance of our *PTPP* scheme in terms of the validity and efficiency.

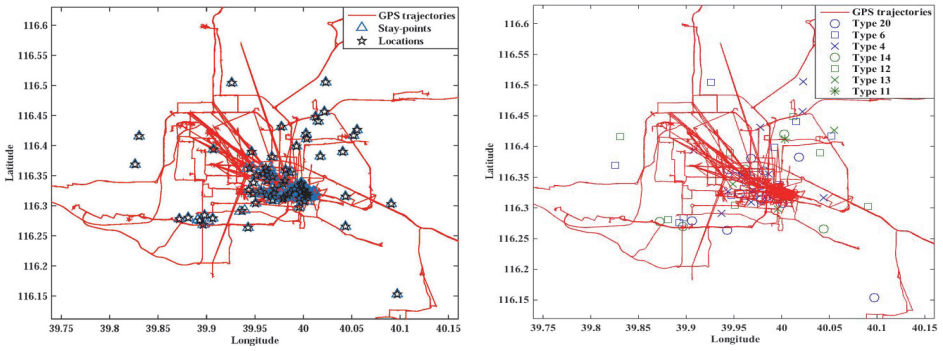
6.1 Datasets and Experimental Setup

We use two real-world datasets in this paper. GeoLife datasets [28] has recorded the GPS trajectory of 182 users with 18670 trajectories in five years (from 2008/10 to 2012/8), which not only includes the daily activity (*e.g.* going home, working, *etc.*), but also includes the recreational activity (*e.g.* shopping, traveling, eating, sports, *etc.*). Most of the data in GeoLife datasets lies in Beijing and few of them are in Europe or USA. POI datasets includes the location information for all kinds of POIs in Beijing [29]. As shown in Table 3, the raw POI datasets can be classified into 20 types.

Table 3. 20 Types of raw POI datasets.

Type	Name	Type	Name
1	Food & Beverages Service	11	Motorcycle Service
2	Road Ancillary Facilities	12	Car Service
3	Place Address Information	13	Car Maintenance
4	Scenic Spot	14	Car Sales
5	Public Facilities	15	Commercial Housing
6	Company	16	Life Service
7	Shopping Service	17	Sports Leisure Service
8	Transportation Service	18	Health Care Service
9	Financial Insurance Service	19	Governments Organizations
10	Education Culture Service	20	Accommodation Service

All experiments are conducted on a computer with Intel i7-3770 3.40 GHz CPU and 4 GB RAM, running 64-bit Windows 7 OS. In our previous experiments [24], we have done some works including stay-point extraction, location extraction and the semantic description of location. Fig. 3 shows the generated preference information of one user.



(a) Location anonymous zone. (b) Semantical description of locations.

Fig. 3. The generated preference information of one user.

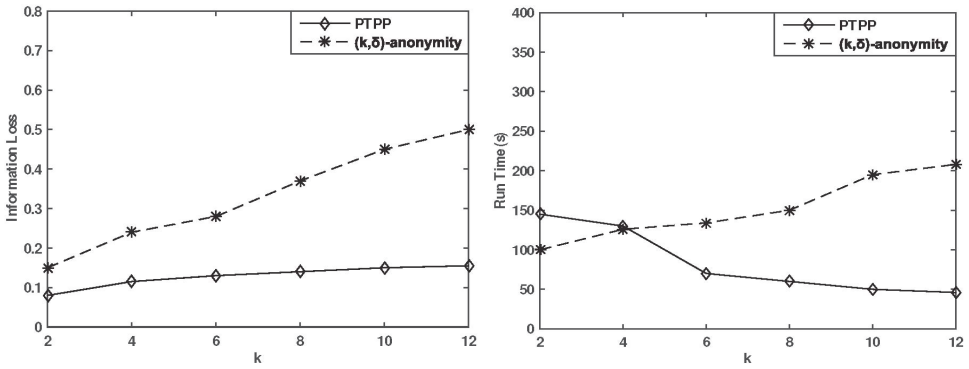
6.2 Data Utility and Efficient Analysis

We compare our PTPP scheme with (k, δ)-anonymity [30] on data utility and efficiency. First, the pre-processing is necessary for (k, δ)-anonymity because it only works in the condition that the time interval of each trajectory is the same. Then, the trajectories are clustered into different clusters. Finally, they are transformed into a (k, δ)-anonymity set.

In order to analyze the data utility of proposed PTPP scheme and (k, δ)-anonymity, we use information loss during the stage of trajectory privacy-preserving. Refer to [31], the information loss can be computed by Eq. (15).

$$Loss_{avg} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(1-1\ / area\left(zone\left(\mathbb{Z}_i, t_j\right)\right)\right)+\sum_{m=1}^{q} L_m}{|T|}, \tag{15}$$

where $area(zone(\mathbb{Z}_i, t_j))$ means the area size of zone $\mathbb{Z}$ at time t_j, L_m means the deleted locations, $|T|$ means the total number of locations in trajectory T.



(a) Information loss. (b) Run time.

Fig. 4. Comparison of data utility and efficiency.

From the result of Fig. 4 (a), we can see that the information loss of PTPP is lesser than (k, δ)-anonymity. The reason is that (k, δ)-anonymity method fails to consider the user preference and privacy risk level in trajectory privacy-preserving. For (k, δ)-anonymity, all the locations are hidden by a unified standard. However, for our PTPP method, the locations can be adaptively hidden according to user preference and background information.

In order to analyze the efficiency of proposed PTPP scheme and (k, δ)-anonymity, we use run time during the stage of trajectory privacy-preserving. From the result of Fig. 4 (b), we can see that the run time of PTPP is shorter than (k, δ)-anonymity when $k \geq 4$. At first, the run time of PTPP is longer than (k, δ)-anonymity. The reason is that PTPP scheme needs to extract stay-points, locations and movement patterns of users at the beginning of trajectory privacy-preserving. However, after the process of user preference modeling, the run time of PTPP scheme is reduced as the value of k increases.

7. CONCLUSIONS

In this paper, we focus on the problem of personalized trajectory privacy-preserving, which considers user preference and background information in trajectory anonymization. We model a user preference to extract stay-points, locations and movement patterns of each user. The proposed preference-aware trajectory anonymization algorithm adaptively selects location privacy-preserving methods by considering user familiarity and location popularity. We conduct a scalable experiment over real-world GPS datasets and POI datasets. The experimental results show that our proposed PTPP method outperforms the existing method in terms of data utility and efficiency.

As a future work, we will refine our trajectory anonymization method by considering distance, velocity, direction, *etc.* Also, we will extend our studies based on other LBSN datasets (*e.g.*, Foursquare, Gowalla, *etc.*) to verify the feasibility.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61522103, 61372112, in part by the joint advanced research foundation of CETC (No. 20166141B08010102).

REFERENCES

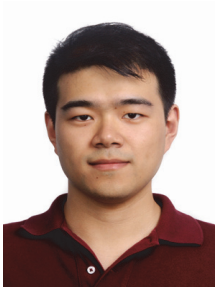
1. Y. Liu, H. Wu, Y. Xia, Y. Wang, F. Li, and P. Yang, "Optimal online data dissemination for resource constrained mobile opportunistic networks," *IEEE Transactions on Vehicular Technology*, Vol. 66, 2017, pp. 5301-5315.
2. C. Xu, Z. Li, J. Li, H. Zhang, and G. M. Muntean, "Cross-layer fairness-driven concurrent multipath video delivery over heterogeneous wireless networks," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 25, 2015, pp. 1175-1189.
3. C. Xu, S. Jia, L. Zhong, and G.-M. Munteant, "Socially aware mobile peer-to-peer communications for community multimedia streaming services," *IEEE Communications Magazine*, Vol. 53, 2015, pp. 150-156.
4. C. Xu, S. Jia, M. Wang, and L. Zhong, "Performance-aware mobile community-based VoD streaming over vehicular ad hoc networks," *IEEE Transactions on Vehicular Technology*, Vol. 64, 2015, pp. 1201-1217.
5. C. Xu, S. Jia, L. Zhong, and H. Zhang, "Ant-inspired mini-community-based solution for video-on-demand services in wireless mobile networks," *IEEE Transactions on Broadcasting*, Vol. 60, 2014, pp. 322-335.
6. X. Zhu, H. Chi, S. Jiang, X. Lei, and H. Li, "Using dynamic pseudo-IDs to protect privacy in location-based services," in *Proceedings of IEEE International Conference on Communications*, 2014, pp. 2307-2312.
7. G. P. Corser, H. Fu, and A. Banihani, "Evaluating location privacy in vehicular communications and applications," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, 2016, pp. 2658-2667.

8. R. Hwang, Y. Hsueh, and H. Chuang, "A novel time-obfuscated algorithm for trajectory privacy protection," *IEEE Transactions on Services Computing*, Vol. 7, 2014, pp. 126-138.
9. L. Zhu, C. Xu, J. Guan, Y. Liu, and H. Zhang, "A preference-aware trajectory privacy-preserving scheme in location-based social networks," in *Proceedings of IEEE International Conference on Computer Communications Workshops*, 2017, pp. 820-825.
10. H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location based services," in *Proceedings of the 21st International Conference on Data Engineering Workshops*, 2005, pp. 88-97.
11. M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu, "Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services," in *Proceedings of IEEE 24th International Conference on Data Engineering*, 2008, pp. 366-375.
12. X. Liu, K. Liu, L. Guo, X. Li, and Y. Fang, "A game-theoretic approach for achieving k-anonymity in location based services," in *Proceedings of IEEE INFOCOM*, 2013, pp. 2985-2993.
13. T. H. You, W. C. Peng, and W. C. Lee, "Protecting moving trajectories with dummies," in *Proceedings of International Conference on Mobile Data Management*, 2007, pp. 278-282.
14. R. Kato, M. Iwata, T. Hara, A. Suzuki, X. Xie, Y. Arase, and S. Nishio, "A dummy-based anonymization method based on user trajectory with pauses," in *Proceedings of the 20th ACM International Conference on Advances in Geographic Information Systems*, 2012, pp. 249-258.
15. S. L. K-anonymity, "A model for protecting privacy," *International Journal on Uncertainty*, *Fuzziness and Knowledge-based Systems*, Vol. 10, 2002, pp. 557-570.
16. B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proceedings of the 25th IEEE International Conference on Distributed Computing Systems*, 2005, pp. 620-629.
17. P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, 2007, pp. 1719-1733.
18. S. Gao, J. Ma, W. Shi, G. Zhan, and C. Sun, "TrPF: A trajectory privacy-preserving framework for participatory sensing," *IEEE Transactions on Information Forensics and Security*, Vol. 8, 2013, pp. 874-887.
19. P. I. Han and H. P. Tsai, "Sst: Privacy preserving for semantic trajectories," in *Proceedings of the 16th IEEE International Conference on Mobile Data Management*, Vol. 2, 2015, pp. 80-85.
20. H. Shin, J. Vaidya, V. Atluri, and S. Choi, "Ensuring privacy and security for LBS through trajectory partitioning," in *Proceedings of the 11th International Conference on Mobile Data Management*, 2010, pp. 224-226.
21. M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proceedings of the 9th International Conference on Mobile Data Management*, 2008, pp. 65-72.
22. M. Gruteser and X. Liu, "Protecting privacy in continuous location-tracking applications," *IEEE Security and Privacy*, Vol. 2, 2004, pp. 28-34.

23. R. Chen, B. C. Fung, N. Mohammed, B. C. Desai, and K. Wang,“ Privacy-preserving trajectory data publishing by local suppression,” *Information Sciences*, Vol. 231, 2013, pp. 83-97.
24. L. Zhu, C. Xu, J. Guan, and S. Yang, “Finding top-*k* similar users based on trajectory-pattern model for personalized service recommendation,” in *Proceedings of IEEE International Conference on Communications Workshops*, 2016, pp. 553-558.
25. L. Zhu, C. Xu, J. Guan, and H. Zhang, “Sem-ppa: A semantical pattern and preference-aware service mining method for personalized point of interest recommendation,” *Journal of Network and Computer Applications*, Vol. 82, 2017, pp. 35-46.
26. J. M. Kleinberg, “Authoritative sources in a hyperlinked environments,” *Journal of the ACM*, Vol. 46, 1999, pp. 604-632.
27. C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, Vol. 27, 1948, pp. 379-423.
28. Y. Zheng, X. Xie, and W. Ma, “GeoLife: A collaborative social networking service among user, location and trajectory,” *IEEE Data(base) Engineering Bulletin*, Vol. 33, 2010, pp. 32-39.
29. http://www.poi86.com/poi/province/131.html, 2014.
30. O. Abul, F. Bonchi, and M. Nanni, “Never walk alone: Uncertainty for anonymity in moving objects databases,” in *Proceedings of the 24th IEEE International Conference on Data Engineering*, 2008, pp. 215-226.
31. R. Yarovoy, F. Bonchi, S. Lakshmanan, and W. H. Wang, “Anonymizing moving objects: How to hide a MOB in a crowd?” in *Proceedings of the 12th International Conference on Extending Database Technology*, 2009, pp. 72-83.

Liang Zhu (朱亮) received Ph.D. degree in Computer Science and Technology from BUPT in October 2017. He is currently a Lecturer with the Institute of Computer and Communication Engineering at Zhengzhou University of Light Industry, Henan, China. His current research interests include mobile social networks, personalized service recommendation, and privacy preserving.

Haiyong Xie (谢海永) received the B.S. degree from University of Science and Technology of China, Hefei, China, in 1997, and the Ph.D. and M.S. degrees in Computer Science from Yale University, in 2008 and 2005, respectively. He is the Director for the Innovation Center, China Academy of Electronics and Information Technology, and a Professor with the School of Computer Science and Engineering, USTC. His research interest includes network traffic engineering, enterprise network traffic optimization, software-defined networking, and future Internet architectures.



Yifeng Liu (刘弋锋) received the Ph.D. degrees in Electronic Engineering from Wuhan University, Wuhan, China, in July 2016. He is currently the team leader of machine intelligence for the Innovation Center, China Academy of Electronics and Information Technology, Beijing, China. His current research interests include around deep learning, machine learning, computer vision, and knowledge engineering.

Jianfeng Guan (关建峰) received the Ph.D. degree in Communications and Information System from Beijing Jiaotong University, Beijing, China, in January 2010. He is currently an Associate Professor with the Institute of Network Technology at Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include around mobile IP, mobile multicast, and next generation internet technology.

Yang Liu (刘杨) received the BE degree in Electrical Engineering and its Automation and the ME degree in Control Theory and Control Engineering from Harbin Engineering University, Harbin, China, in 2008 and 2010, respectively, and the Ph.D. degree in Computer Engineering at the Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, in 2014. He is currently a Lecturer with the Institute of Network Technology at Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include wireless networking and mobile computing. He is a member of the IEEE and ACM.

Yongping Xiong (熊永平) received his Master of Computer Network Security from Harbin Institute of Technology, and Ph.D. of Engineering in Computer Architecture from Institute of Computing Technology, Chinese Academy of Sciences in 2005 and 2010, respectively. Since July 2010, he has been an Associate Professor at State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications. His main research interests cover mobile opportunistic network, distributed system and mobile computing.