IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2017, accepted October 9, 2017, date of publication October 30, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2767641

A Study of Enhancing Privacy for Intelligent Transportation Systems: *k*-Correlation Privacy Model Against Moving Preference Attacks for Location Trajectory Data

PEIPEI SUI[1], XIANXIAN LI[2], AND YAN BAI[3]

[1]School of Management Science and Engineering, Shandong Normal University, Jinan, China
[2]Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, China
[3]Institute of Technology, University of Washington Tacoma, Tacoma, WA, USA

Corresponding author: Xianxian Li (lixx@gxnu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61672176 and Grant 61502111 and in part by the Guangxi Special Project of Science and Technology Base and Talents under Grant AD16380008.

ABSTRACT Internet of Things (IoT) has been widely used in various application domains including smart city, environment monitoring and intelligent transportation systems. Thousands of interconnected IoT devices produce an enormous volume of data termed as big data. However, privacy protection has become one of the biggest problems with the progress of big data. Personal privacy is usually challenged by the development of technology. In this paper, we focus on privacy protection for location trajectory data, which is collected in intelligent transportation system. First, we demonstrate that the moving preference of individuals can be exploited to perform re-identification attacks, which may cause serious damage to the identity privacy of users. To address this re-identification problem, we present a new trajectory anonymity model, in which the degree of correlation between parking locations and individuals is precisely characterized by a concept of Location Frequency-inverse user frequency (LF-IUF, for short). We then propose an anonymizing method to replace parking locations by a k-correlation region. Our method provides a novel anonymity solution for publishing trajectory data, which achieves a better trade off between privacy and utility. Finally, we run a set of experiments on real-world data sets, and demonstrate the effectiveness of our method.

INDEX TERMS Data privacy, location, information security, Internet of Things.

I. INTRODUCTION

The IoT provides new services and applications that can be deployed in smart homes, smart cities, and intelligent transportation systems. A large amount of data can be collected from multiple IoT devices, which can be utilized for big data analysis. However, in many such applications the data collected by IoT is sensitive and must be kept private and secure, such as patient data in healthcare and location data in transportation. The concept of privacy in different domains may differ, but in general, it should protect user's personally identifiable information and keep a certain degree of anonymity, unlink-ability and data secrecy. Existing techniques for protecting sensitive data in IoT have mainly focused on securing the communication channel, as well as user authentication and authorization. Little work has been done to protect sensitive data after they are collected [1]. The disclosure of such data may create opportunities for criminal activity, or result in serious harm to application users. To protect such sensitive data, we should design anonymity methods to ensure IoT big data privacy and security. In this paper, we take location trajectory data in intelligent transportation systems as an example to demonstrate the privacy protection of trajectory data.

Studies on vehicle trajectory datasets have always been important topics for many location-based applications such as mobile navigation system, urban traffic analysis. Such datasets are full of value and can be used in many fields. For example, analyzing trajectories of passengers in an area can help people make commercial decisions, such as where to build a restaurant; analyzing trajectories of vehicles in a city may help government to optimize traffic management systems. However, personal location information is highly

sensitive, since a location trace not only is a set of positions on a map, but also tells much about our habits, interests, activities and relationships.

Therefore, location trajectory data need to be anonymized before being published. As trajectory contains abundant spatiotemporal information, simply removing identifiable attributes is insufficient to protect the privacy of individuals.

Huo *et al.* [2] explained the ways of protecting parking locations information in order to prevent sensitive information leakage on location trajectories. A target user will be re-identified if the parking location information is related enough to an individual. The parking location can play the role of a quasi-identifier (or QID) to re-identify the user. To avoid the issue of re-identification of a user, generalizing parking locations was adopted. However, the generalization of parking locations could lead to information loss. It was, therefore, suggested to modestly adopt according to the privacy requirements of users in order to minimize information loss. It is not necessary to incorporate all parking locations into this privacy strategy, such as generalization.

The privacy risk of parking location can be characterized with a probability of re-identifying a trajectory based on parking locations. Assuming that an adversary has been able to observe the victim's movements during a certain time and try to infer the identity of the victim. The success probability of identity attack mainly depends on the correlation between parking locations and users. In previous works [25], [26], the correlation between parking locations and users were evaluated by the frequency of location in history location trajectories of users. However, it is not precise enough; it ignores the effect of inverse-frequency. That is, a higher frequency of a parking location's occurring in a user's history trajectories does not imply a higher probability of re-identification, and a lower frequency does not mean the lower privacy risk of the parking location. We need a more rational approach to evaluate the privacy risk.

In this paper, we first learn each user's moving preference from the history trajectory datasets, and then quantify the correlations between locations and users. That is, there is a correlation value to describe the relationship between each parking location and a user. Next, we present a re-identification attack based on the user's moving preference. Consequently, a privacy-protection model is proposed to avoid the over-protected problem in privacy protection of location trajectory publishing.

Our contributions are as follows:

We study re-identification attacks based on history location trajectory information, namely moving preference attack. To the best of our knowledge, this topic has not been well studied by prior works. We propose a new concept, called Location Frequency-Inverse User Frequency, to evaluate the information leakage of parking locations.

- A k-correlation privacy model is proposed for resisting the moving preference attack. This model more precisely measures the privacy risk of parking locations, as compared with several existing related models. We also develop a novel anonymity technique to realize the k-correlation model. Our technique not only enhances the identity privacy of trajectories, but also decreases the number of generalized parking locations, consequently, reducing information loss.
- We evaluate the performance of our methods on a real world dataset. The results show that our method achieves a better privacy-utility trade-off than existing related models, including k-anonymity and GridPartition [2].

The rest of the paper is organized as follows. Section II discusses the related works on privacy protection of trajectory data publishing. In Section III, some background information related to our methods and notions are given. Section VI states the problem and describes our methods. The experimental results on a real world dataset are presented in Section V. Finally, Section IV provides the conclusion.

II. RELATED WORK

In this section, we briefly describe the most related topics in trajectory privacy protection. Interested readers may find more details in recent surveys and tutorials [3], [4].

A. PRIVACY RISK

The spatial and temporal attributes of a trajectory can be considered as powerful *QIDs* that can be linked to other kinds of physical data objects. Researchers illustrated trajectory risk by inferring a user's home address through location trajectory information. Liao *et al.* [5] used a time and location-sensitive clustering algorithm to find a user's frequent destinations based on higher-level machine learning about a user's habits. They attempted to infer the user's household geographic location. Hoh *et al.* [6] used a database of week-long GPS trajectories from 239 drivers in the Detroit, MI, USA area. They designed a clustering-based identification algorithm to find plausible home addresses of about 85% of the 65 drivers. John [7] collected 172 people's GPS data, and analyzed data using four heuristic algorithms which were the *Last Destination*, *Weighted Median*, *Largest Cluster*, and *Best Time* Algorithms, to compute the coordinates of each driver's home address.

Shokri *et al.* [28] represented the user profiles using the hidden Markov model, and formalized the adversary's performance. Basically, they computed a matching probability between pseudonym traces and user profiles using the classical forward-backward algorithm. The matching probability represents the likelihood that a particular set of traces corresponding to a specific user. De Montjoye *et al.* [29] showed that if an individual has a unique pattern in the anonymized dataset then it is enough to identify him even if the dataset does not contain personal information, such as name, age and address. Hua *et al.* [30] presented an attack based on a semi-supervised learning approach to infer the riding trajectory of the user.

B. PRIVACY MODEL

k-anonymity is often used in both location-based services privacy protection and trajectory publishing privacy protection. Abul *et al.* [8] proposed an extended concept of k-anonymity

based on the imprecision of sampling and positioning systems, named (k, δ)-anonymity, δrepresents the possible location imprecision. Based on space translation, the general idea is to modify the paths of location trajectories so that k different trajectories co-exist in a cylinder of the radius δ. However, the imprecision assumption may not hold in some sources of trajectory data, such as transit data, RFID data and purchase records. Yarovoy *et al.* [9] presented another notion of k-anonymity based on the assumption that different trajectories may have different *QIDs*. Specifically, they considered timestamps as *QIDs*, and assumed that adversaries conduct privacy attacks based on an attack graph. Josep *et al.* [10] introduced a similar micro aggregation approach which has been successfully used in micro-data anonymization to achieve privacy protection in trajectory data publication. They first clustered trajectories into clusters the size of at least k based on their similarities and then replaced a cluster of trajectories with synthetic data that preserved all the visited locations and a number of original trajectories. Although they offered better utility than (k, δ)-anonymity, the information loss is still large.

Some works consider that the potential *QID*s are more complex in trajectory databases. Mohammed *et al.* [11] limited adversaries' background knowledge by a parameter L, and presented a *LKC*-privacy model, where L is the maximum length of background knowledge. Giorgos *et al.* [12] proposed a distance-based generalization to achieve k^m-anonymity for location trajectory data. Similar to the *LKC*-privacy model, it does not require detailed knowledge of *QID*s, or a distinction between sensitive and non-sensitive information, prior to data publishing.

Another privacy protection method is suppression. It simply deletes sensitive locations, and thus protects them in trajectory data publishing. Terrovitis *et al.* [13] assumed that each adversary would possess different portions of users' trajectories and that the data publisher was aware of the adversaries' knowledge. They proposed a method that iteratively suppressed some trajectory segments until a probabilistic constraint of disclosing whole trajectories was satisfied. However, if too many trajectory segments are suppressed, it will cause much information loss. Subsequently, Chen *et al.* [14] first proposed a local suppression method to solve the problem. However, it often leads to high information loss and ineffective data mining.

Differential privacy has been also widely used to protect the privacy of individual participants while providing useful statistical information about the whole population. Chen *et al.* [15] was among the first to connect trajectory data publishing with differential privacy, and proposed a data dependent sanitization mechanism by building a noisy prefix tree under a Laplace mechanism. Their recent work [16] developed a robust sampling-based framework to systematically explore the dependencies among all attributes, and subsequently, build a dependency graph. It preserves the joint distribution of high-dimensional data under differential privacy.

Some other works related to privacy are based on road network, and semantic trajectory privacy [17], [18]. Monreale *et al.* [17] claimed that semantic trajectory poses important privacy threats. They defined an attack model of semantic trajectory linking, together with a privacy notion, c-safety, based on the generalization taxonomy of visited places. It provides an upper bound c in relation to the probability of inferring that a given person, observed in a sequence of non-sensitive places, has also stopped in any sensitive locations. Yigitoglu *et al.* [18] presented an approach to protect sensitive semantic positions in an urban setting, which extends the semantic location cloaking model to protect against velocity-based linkage attacks.

In literatures, most anonymization approaches in a spatiotemporal context are based on randomization techniques, space translations, and suppression of various portions of a trajectory [19]. They anonymize trajectories under a unified standard, which may lead to serious information loss. Our work is inspired by [25] and [26]. Zang *et al.* [25] conducted a large scale study on the risk of re-identification attacks with call records data, and considered the "top N" locations visited by users. They showed that the number N of top preferential locations determines the power of an adversary and the safety of a user's privacy. The more top locations of a victim an adversary knows, the easier it is to identify the victim. Gambs *et al.* [26] focused on de-anonymization attacks, by which adversaries try to infer the identity of an individual from a set of mobility traces. They modeled an individual's mobility using a mobility Markov chain to perform the de-anonymization attack. The experiment results showed that a de-anonymization attack can re-identify individuals whose movements are contained in an anonymous dataset providing some mobility trajectories that the adversary can use as background knowledge. However, we consider this problem from a different perspective. Instead of focusing on re-identification and de-anonymized attacks, we considered the efficiency of privacy protection method against those attacks.

III. PROBLEM FORMULATION

Spaccapietra *et al.* [20] proposed the first model that treats trajectories of moving objects as a spatiotemporal concept. They conceptualized a trajectory as a space-time evolution of an object to reach a certain goal. In this section, we will present some definitions relevant to trajectory privacy and illustrate our problem's definition.

A. PRELIMINARY

Definition 1 (Trajectory): By ordering the GPS points from a moving object by time, we can obtain a trajectory. A trajectory of a moving object u is a sequence of time-stamped points $tr =< u, p_0, p_1, \ldots, p_n >$, where $p_i = (x, y, t)$, t is a timestamp and (x, y) is two-dimension coordinate, which represents the latitude and longitude of points.

We use $Tr(u)$ to denote the set of trajectories belongs to u in a trajectory dataset. We also use $Dist(p_i, p_j)$ to stand for the

geospatial distance between two GPS points p_i and p_j, and $Int(p_i, p_j) = |p_i.t - p_j.t|$ for the time interval when a moving object moves from p_i to p_j. Let T_r and D_r be the time and distance thresholds respectively. The parking point is defined as follows.

Definition 2 (Parking Point): A parking point pp is a 5-tuple. $pp =< u, x, y, t_a, t_l >$, where (x, y) is the location coordinate of a parking point, t_a is the arriving time and t_l is the departure time.

Parking points are generated by a user staying over a certain amount of time T_r within a distance threshold of D_r. In a trajectory $< u, p_0, p_1, \ldots, p_n >$, if there is a subsequence $< p_i, p_{i+1}, \ldots, p_j >$, such that $Dist(p_i, p_k) \leq D_r$ for any $k(i < k \leq j)$, $Dist(p_i, p_{j+1}) > D_r$ and $Int(p_i, p_j) \geq T_r$, then we can obtain a parking point $pp=< u, x, y, p_i.t, p_j.t >$, where $x = \sum_{k=i}^{k=j} p_k.x/(j-i+1)$, $y = \sum_{k=i}^{k=j} p_k.y/(j-i+1)$.

Parking points of a real-world location may have different coordinates. That is, although multiple users visit the same location, such as a shopping mall, the parking points we extract are different. Besides, the imprecision of GPS devices may also result in different parking points for a real-world location. So we need to define the notion of parking location. To facilitate defining parking location, we use $pp \leftarrow tr$ to denote that the parking point pp is extracted from a moving object u's trajectory tr.

Definition 3 (Parking Location): A parking location l is a geographic place constructed by a set of parking points $\{pp_i\}$, denoted by the formula $l = \mathscr{C}\{pp_i\}$. In this case, we also say the parking location l is generated by $\{pp_i\}$, denoted by $pp_i \in_C l$.

B. FORMULATION AND PROBLEM DEFINITION

Definition 4: Suppose$U$ is a set of users and L is a set of parking locations extracted from a dataset D. A user parking frequency function f is defined as follows.

$$f(u, l) = |\{pp_i | pp_i \in_C l, pp_i \in P_u\}|, \tag{1}$$

where $u \in U$, $l \in L$ and $P_u = \{pp_i | pp_i \leftarrow tr$ and $tr \in Tr(u)\}$.

Definition 5 (Location Frequency): Let u_i be a user, and l_j be a parking location. A location frequency function LF is defined as follows.

$$LF(u_i, l_j) = \frac{f(u_i, l_j)}{\sum_{l_q \in L} f(u_i, l_q)} \tag{2}$$

For the re-identification attack, LF is not precise enough to evaluate the privacy risk of a parking location. For example, suppose l is a public place. The higher visited frequency by u does not imply that l is more sensitive; many other users might visit this place. We, thus, use inverse user frequency to test whether a parking location is common or rare across all users. The definition of inverse user frequency is shown in Definition 6.

Definition 6 (Inverse User Frequency): The inverse user frequency function IUF is defined as follows.

$$IUF(l_j, U) = \log \frac{|U|}{\left|\left\{u_i | u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|}, \tag{3}$$

where U is a set of users, and l_j is a parking location. For the convenience of calculation the log function is used, that reduces the absolute value of IUF. By combining LF and IUF, we propose the function LF-IUF ($Location\ Frequency$–$Inverse\ User\ Frequency$) to evaluate the sensitivity of a parking location for re-identifying a user. Definition 7 describes LF-IUF in detail.

Definition 7 (LF-IUF): LF-IUF is defined as:

$$LF - IUF(u_i, l_j, U) = LF(u_i, l_j) \times IUF(l_j, U). \tag{4}$$

LF–IUF is a numerical statistic that is intended to reflect how important a location l is for a user u in a space. The value of LF weighs l's importance for u without considering other users, and the value of IUF reflects the distribution of location l among the entire users U. A high LF–IUF value occurs when the location frequency of a given user is high and the user frequency of the location in the whole collection of trajectories is low; hence, LF–IUF tends to filter out common locations shared by all users.

Now, we give an example to evaluate correlation degrees of different locations. Assume that we have a parking location dataset as described in Table 1. It contains five users and four parking locations.

TABLE 1. Parking times of users.

Users	l_1	l_2	l_3	l_4
u_1	3	0	0	0
u_2	2	2	0	0
u_3	0	0	3	1
u_4	0	0	2	0
u_5	2	0	0	1

We can compute LF according to (2) and IUF using (3). We then obtain the IF-IUF results as shown in Table 2.

TABLE 2. Values of IF-IUF.

Users	*LF*				*LF-IUF*			
	l_1	l_2	l_3	l_4	l_1	l_2	l_3	l_4
u_1	1	0	0	0	0.511	0	0	0
u_2	1/2	1/2	0	0	0.256	0.805	0	0
u_3	0	0	3/4	1/4	0	0	0.687	0.229
u_4	0	0	1	0	0	0	0.916	0
u_5	2/3	0	0	1/3	0.341	0	0	0.305

We notice that u_1 has only one parking location l_1 and the $LF(u_1, l_1)$ has reached the maximum value of 1, but l_1 is also visited by u_2 and u_5, so the LF-$IUF(u_1, l_1, U)$ is offset some

correlation degree by IUF(u_1, l_1). We observe that u_3 has two parking locations l_3 and l_4 with the same value of IUF, while LF(u_3, l_3) > LF(u_3, l_4). So, the association degree LF-IUF of location l_3 is higher than that of l_4 to user u_3.

Based on above-mentioned analysis, the correlation between locations and individuals can be evaluated by *LF-IUF*, and then the value of *LF-IUF* could reflect individual's moving preference. In this paper, we focus on the re-identification attack based on the moving preference of individuals called the *Moving Preference Attack*, by which adversaries try to infer the identity of some location trajectories in published datasets. In this paper, we use *LF–IUF* (u_i, l_j, U) to describe the correlation between parking location l_j and individual u_i, where U is the set of moving objects. For a published trajectory dataset D, we define a moving preference attack as follows.

Definition 8 (Moving Preference Attack): We assume that adversaries know the parking preference of the victim u_i and can access the published trajectory dataset D. The parking preference of u_i can be learnt by the adversary's observation or from history location trajectories in the past. Then, the objective of the adversary is to de-anonymize dataset D by linking it to the corresponding individual u_i.

Our main goal is to provide a model and related approach to generate an anonymous trajectory dataset, which guarantees that the probability of attackers' being able to infer the identity of a user is less than a pre-defined threshold. For privacy protection, it is desirable to minimize the correlation between a user and a parking location. The set of candidate trajectories corresponding to a given parking location should be as large as possible. We propose a concept of k-correlation as a compromise.

Definition 9 (k-Correlation): A parking location (or a region consisting of several locations) R is k-correlation for a moving object u_i, if

$$LF-IUF(u_i, R, U) \leq \frac{1}{|\{l|tr \in tr(u_i), pp \leftarrow tr, pp \in_C l\}|} \times \log(\frac{|U|}{k}), \quad pp \in R \text{ and} \tag{5}$$

where u_i denotes a moving object, and pp is a parking point of a trajectorytr, $|\{l|tr \in tr(u_i), pp \leftarrow tr, pp \in_C l\}|$ is the total number of parking locations from u_i.

To capture the information loss, we adopt the reduction in the probability with which people can accurately determine the position of an object in [9] as our information metric.

Definition 10 (Information Loss): Given an anonymized version D^* of a trajectory dataset D, the information loss is measured by

$$IL_{avg} = \frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(1 - 1/area(region(tr_i, l_j)))}{N \times M}, \tag{6}$$

where $area(region\ (tr_i, l_j))$ represents the area size of the corresponding *k-correlation* region of u_i when tr_i contains l_j. We define area size as the number of cells a region covers. N denotes the total trajectories in D, and M denotes the number of different "top m" parking locations of history trajectories. N^*M represents the number of locations that anonymized in D^*. So, IL_{avg} ranges from 0 to 1.

Based on the definitions introduced in the above, our problem definition is as follows.

Definition 11 (Problem Definition): Given a location trajectory dataset D and a protection threshold k, k-correlation privacy model generates a version D^* of D such that all parking locations in D^* satisfy k-correlation.

IV. ANONYMIZED METHOD AGAINST MOVING PREFERENCE ATTACKS

In this section, we describe an anonymity preserving method to protect location trajectory data against moving preference attacks based on k-correlation. It first extracts the parking locations of each individual. A parking location is strongly correlated to an individual when it allows inference of the individual's identifier. To avoid this re-identification attack, we should point out which locations are closely correlated to individuals and how strong the association is. Then, for every parking location that is correlated to a user, we will replace it with a k-correlation region to anonymize the trajectory.

A. EXTRACTING PARKING LOCATION

Users usually choose a fixed range to drive and fixed location to stop. So we can learn the parking habits of drivers from their history trajectories. The first step is to detect parking locations from the GPS data. Here, we extract every parking location from the vehicle GPS dataset according to Definition 3. We set the time interval T_r as 15 minutes and the distance threshold D_r as 40 meters. Interested readers can find more detailed information about parking location extraction in [2], [20], and [21].

B. COMPUTING THE LF-IUF

As mentioned in [26], the mobility of an individual can act as a signature, and plays the role of a quasi-identifier. If an adversary learns a victim's mobility signature from history trajectories, he can identify the victim by finding a matched signature in the anonymized dataset. To address this re-identification attack, we need find the most likely signature (*QID*) of victims, and then increase the size of anonymized set by adding more of the same *QID*s. In our work, we first find the closely correlated locations of each individual without considering others, called personal-correlated locations. We then adopt *LF-IUF* to evaluate the degree of association between parking locations and individuals.

As shown in Fig. 1, if a user u stopped at a location l, there will be a link from the user to the location. A user can park at several different locations, and a location can be parked at by many different users. In Fig.1, gray nodes denote moving objects, black nodes stand for parking locations, and weight f_{ij} on an edge represents the frequency of a user u_i parking at the location l_j. The parking frequency f_{ij} can be learned from a user's history trajectories [22].

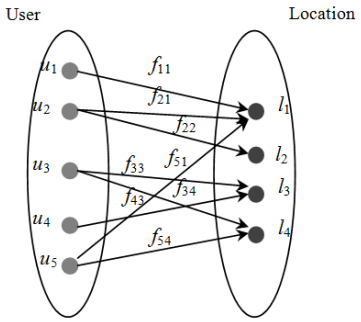


FIGURE 1. Users and parking locations.

Given a collection of user's parking locations extracted from a history trajectory dataset, we can build a matrix M, in which an entry f_{ij} stands for the frequency that u_i has parked at location l_j, $1 \leqslant i \leqslant |U|$, $1 \leqslant j \leqslant |L|$. The matrix M can be represented as follows:

$$M = \begin{bmatrix} f_{11} & f_{11} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{bmatrix} \tag{7}$$

The raw frequency f_{ij} in a user's history trajectories is the number of times that u_i has been ped at l_j. To prevent a bias towards more trips, we select an augmented frequency to compute $LF(u_i, l_j)$, the raw frequency divided by the sum raw frequency of all parking locations of u_i is described by (8).

$$LF(u_i, l_j) = \frac{f_{ij}}{\sum_{q=1}^{q=m} f_{iq}} \tag{8}$$

The inverse user frequency is a logarithmically scaled fraction of the users that stopped at the location. It is obtained by dividing the total number of users, n, by the number of users parking at the location, and then taking the logarithm of that quotient, as described by (9).

$$IUF(l_j, U) = \log \frac{n}{\left|\left\{u_i | u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|}, \tag{9}$$

where $\left|\left\{u_i | u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|$ denotes the number of users that parked at location l_j, n is the number of users in U. *LF-IUF* is then calculated using (10):

$$LF - IUF(u_i, l_j, U) = \frac{f_{ij}}{\sum_{q=1}^{q=m} f_{iq}} \times \log \frac{n}{\left|\left\{u_i | u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|}. \tag{10}$$

It is noted that the above formula consists of two components. One is the location frequency (*LF*) and the other is the inverse user frequency (*IUF*). The *LF* is calculated as the frequency of location l_j divided by the frequency of all the locations from a user u_i's history trajectories. Its value reflects the correlation between location l_j and user u_i without considering other users. On the contrary, the value of *IUF* shows the influence of other users in the database. As a location is visited by more users, the ratio inside the logarithm approaches 1, consequently, bringing down the value of *IUF* closer to 0. The *LF-IUF*(u_i, l_j, U) is proportional to the visited frequency of u_i and is inversely proportional to the number of users parking at l_j. A high *LF–IUF* value occurs when the location frequency of a given user is high and the user frequency of the location in the whole collection of trajectories is low. So locations with high values of *LF-IUF* are considered to be generalized before published.

C. TRAJECTORY ANONYMIZATION

As in [26], we assume that an adversary observed the history of movements made by some individuals during a non-negligible amount of time, which can be treated as background knowledge of an adversary. Later, the adversary accesses a different trajectory dataset containing mobility traces of the individuals observed previously. The aim of the adversary is to re-identify this dataset by linking it to the corresponding individuals contained in their background knowledge.

In process of trajectory anonymization, we attempt to prevent adversaries from inferring individuals from the anonymity dataset, which means the value of *LF-IUF* of personal-correlated location has to be smaller than a sensitivity threshold. According to the above analysis, our object is to reduce the correlation between users and parking locations. There are mainly four approaches [3] to protect trajectory data publication privacy. Firstly, the clustering-based approach adopts the uncertainty of trajectory data to group k trajectories within the same time period to form a k-anonymized aggregate trajectory. Secondly, the generalization-based approach picks atomic points from the group and rebuilds trajectories based on these points. Thirdly, the suppression-based approach deletes locations iteratively until the privacy constraint is met. Lastly, the grid-based approach aims to construct a grid on a system space and partition the grid based on the privacy requirements. The grid-based approach [2] is a simple and effective method, which divides the space map into several grid cells. It supports most location trajectory queries in data mining. In this paper, we propose a *TRAMP* Algorithm (trajectory-anonymity against moving preference) to anonymize trajectories by combining spatial cloaking on a grid. We divide the space map into grid cells on-demand. On one hand, a coarse grid may have a very low accuracy because the area covered by each grid node is too large. A fine grid needs more storage and computing resources.

Next, we identify personal-correlated locations of each individual. If the value of *LF-IUF* (u_i, l_j,) is less than the sensitivity threshold ∂, then we publish this parking location directly. Otherwise, we apply a greedy heuristic algorithm to select one neighbouring l_{neig} with the minimum value of *LF-IUF* to merge, replace the value of *LF-IUF*(u_i, l_j,) by *LF-IUF*(u_i, $\{l_j, l_{neig}\}$). We repeat the merging process until the privacy requirement is met.

The proposed trajectory anonymized algorithm based on *LF-IUF* is described as follows.

Algorithm 1 Trajectory-Anonymity Against Moving Preference Algorithm

Input: trajectory dataset *TrajD;* personal related demands k history trajectories *TrajH*

Output: the published trajecotry dataset $TrajD^*$

1: $TrajD^* \leftarrow TrajD$
2: Extracting parking locations L_H for each user from *TrajH*
// computing frequency matrix M
3: for each $u_i \epsilon U$ and each $l_j \epsilon L_H$ do
4: $M[i][j] \leftarrow f(u_i, l_j)$
5: Extracting parking locations L for each user from *TrajD*

// generate k-correlation region R
6: for each $u_i \epsilon U$ and each $l_j \epsilon L$ do
7: while $LF\text{-}IUF(u_i, l_j)) > (1/m) \times \log(|U|/k)$ do
8: select a neighbor l_k with the minimum *LF-IUF* value to merged
9: replace l_j by the merged region $R = \{l_j, l_k\}$
10: update *LF-IUF*
// generalize trajectories
11. for each t_i in *TrajD*
12: if t_i's point is in the related region then
13: replaced it by its k-correlation region R in $TrajD^*$
14: else
15: preserved*in* $TrajD^*$
16: Return $TrajD^*$

Three inputs of Algorithm 1 are trajectory dataset to be published*TrajD*, a correlation parameter k, and a history trajectory dataset *TrajH*. The algorithm starts with extracting parking locations from a history of trajectories; and then defines a frequency matrix M. The entry $M[i][j]$ denotes a frequency of user u_i parking at location l_j. According to (10), we can get the value of *LF-IUF*. In line 7, we set the sensitivity threshold ∂ as $(1/m) \times \log(|U|/k)$. Here, m denotes the number of entries which are greater than zero in the ith row of M. It is the number of locations where u_i has been parked, and k is an integer number greater than zero.

We obtain the k-correlation region R for a user u_i in respect to a parking location l_j as follows. For each user in the *TrajD*, the correlation *LF-IUF* between u_i and l_j is calculated in line 6, and k-correlation is tested in line 7. If some parking location violates k-correlation, that is the value of *LF-IUF* is greater thanu_i's threshold ∂, we then pick a neighbor location with a minimum *LF-IUF* amongst all the neighbors to merge. This merging process repeats until the correlation condition is satisfied. We will get a k-correlation region R for u_i at l_j (lines 7-10). At the last step, if a GPS point locates in a k-correlation region, it will be replaced by the region in anonymized trajectory data; otherwise, the point can be published directly.

Note that a location may be generalized to multiple regions in different trajectories. Trajectories are anonymized based on the grid cell map. The original trajectory dataset is set as the input. Each GPS sample is scanned. Personal-correlated locations are replaced by the corresponding k-correlation region. For other points in the GPS sample, the published version is the same as the original one, unless the GPS point is located in the k-correlation region of the same trajectory.

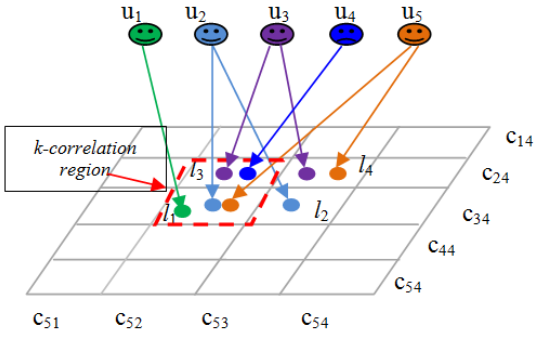


FIGURE 2. Personal-correlated locations.

Figure 2 depicts an example of generalization with $k = 2$. User u_3 has two parking locations l_3 and l_4 located at grid cells c_{22} and c_{23}. We can easily calculate the user's personal correlation degree of l_3 and l_4, the results are shown in Table 2.

Next, we compare $LF\text{-}IUF(u_3, l_3)$ and$LF\text{-}IUF(u_3, l_4)$ with the sensitivity threshold ∂. If $LF - IUF < \partial$, then we publish this parking location directly, otherwise, we perform an estimation based on cell c_{22} to decide whether this region is k-correlation region or not. If not, we merge the proper neighbor cell, and repeatedly merging until we find the k-correlation region of the parking location l_3. In this example, we set $k = 2$, then the threshold $\partial = 1/2 * \log(5/2) = 0.458$, we observe that $LF\text{-}IUF(u_3, l_3) > \partial$ and $LF\text{-}IUF(u_3, l_4) < \partial$, as we calculate, the region composed by $l_3(c_{22})$ and $l_1(c_{32})$ is a k-correlation region of l_3. Therefore, for the trajectory of u_3, l_4 can be published directly, and l_3 should be replaced by the region contained l_3 and l_1.

V. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of our trajectory anonymization method described in Section IV.

A. DATASET

We use a real-world taxi trajectory dataset in our experiments. It contains about 356,000 trajectories of 12,504 taxis. The total number of GPS points is 596 million. The interval of GPS data collection is approximately one minute. All taxis simultaneously report a file every five minutes. Each file records taxi's SIM card number, longitude and latitude of

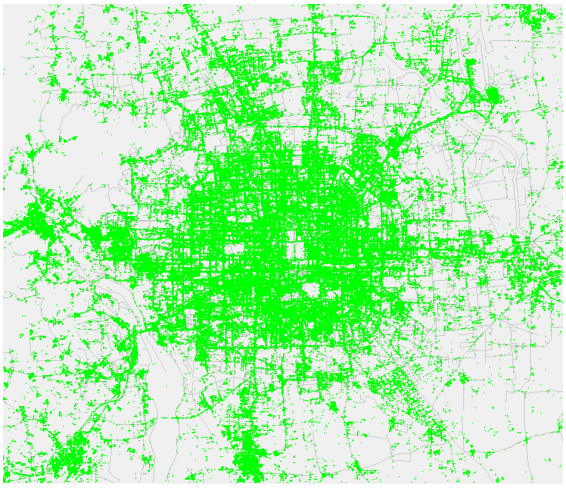


FIGURE 3. Spatial distribution of parking locations.

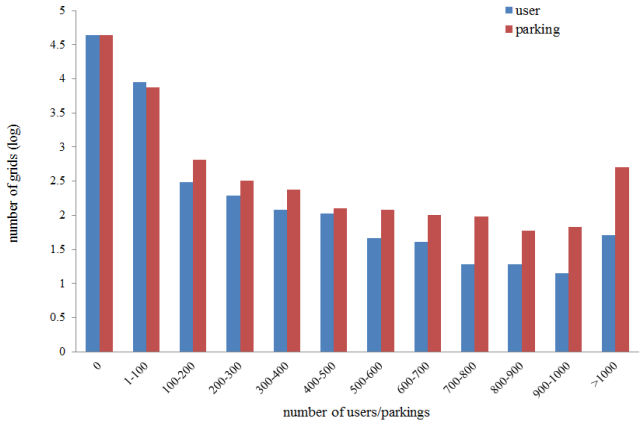


FIGURE 4. Number of grids with different parked frequency.

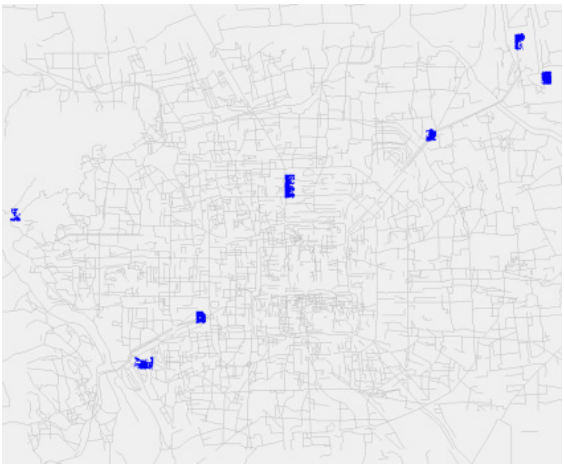


FIGURE 5. Top 10 parking locations.

the vehicle's location, time, speed, direction, current status of the vehicle and other information during the five-minute time interval.

There are 9548 taxis left after data preprocessing. We set the time threshold to 15 minutes and the distance threshold to 40 meters for parking location detection just like [22], which results more than 1.9 million parking locations. Figure 3 shows the spatial distribution of parking locations. Obviously, the distribution of parking locations within the Fifth Ring Road in Beijing is concentrated. They almost cover the whole area. Locations outside the Fifth Ring Road are distributed comparatively sparsely.

To simplify the generalization process of locations, we use a grid-based approach to split the map as aforementioned [2]. The map is divided into 53186 grids; users have stopped at only 9761 grids. Therefore, the parking coverage of grids is only 18%, which suggests that parking locations are relatively concentrated. When analyzing the 9761 grids, we find that the average number of users per grid is 47 during a month, and the average number of parking points per grid is 202 during a month. This suggests that a user might have stopped at a location several times. Figure 4 presents the grid's distribution with differences in the frequencies of users and parking locations. The x-axis shows the number of users (parking times) per grid, and y-axis shows the number of grids in log-scale.

Note that in Fig. 4, there are 8845 grids with the parking frequency between 1 and 100 times, and 7456 grids have been parked at by 1 to100 users. 511 grids have been parked at more than 1000 times, and 51 grids parked at by more than 1000 users. Next, we find the ten most popular grids. Figure 5 depicts the top 10 parking locations (grids). We used Google Maps API to reverse these parking locations to a real-world address. They are located at Deshengmen, Sanyuanqiao, Liuliqiao, Bird Nest, Beijing Capital International Airport, train stations, and subway's terminal stations, respectively. The most popular grid is located at Beijing Capital International Airport, which is the biggest airport of China with its passenger volume at the 2*nd* place worldwide. Bird Nest and Water Cube have been the tourist hot spots after the 2008 Beijing Olympic Games. It is reasonable that there are more taxis waiting for passengers at these locations. It is also reported that there are giant free parking lots at Deshengmen, Sanyuanqiao, Liuliqiao.

This result further illustrates that a user might have parked at a location several times and that different grids have different parking frequencies. The correlation between a user and a location is not only dependent on the user's parking frequency, but also related to the location's popularity. It is reasonable to compute the personal correlation degree using *LF-IUF* as defined in Section III.

B. PRIVACY ANALYSIS

In this section, we discuss the privacy guarantees of our anonymization approach. We formally show that the *TRAMP* Algorithm guarantees that the anonymized datasetD^* is a k-correlation version of D.

In our model, we assume that the background knowledge obtained by adversaries can be modelled as a matrix, whose entry $p_{ij} = LF(u_i, l_j)$ is the probability that u_i stops at location l_j to describe the moving preference of users. We quantify

"how much correlation" between an arbitrary location and a user is by a correlation metric function *LF-IUF*. We select the "top m" locations parked by users to describe users' parking preferences. As declared in [25], it is natural to consider that the top 2 locations are home and work places for majority users, but, more generally, the number m of top preferential locations determines the power of an adversary and the safety of a user's privacy. The more top locations of a victim an adversary knows, the easier it is to identify the victim.

Theorem 1: Given a trajectory dataset D and a correlated threshold k⩾1, Algorithm TRAMP produces a dataset D^* that is a k-correlation version of D.

Proof: For any parking location l_j of a user u_i, there are three possible cases:

(a) Location l_j is not in the "top m" locations of individual u_i, which implies that the mobility behavior of u_i is irrelevant to this location l_j. In our assumption, adversaries take the probability of selecting top m parking locations of an individual as their background knowledge. In this case, location l_j can be released directly without violating the concept of k-correlation.

(b) Location l_j belongs to the "top m" locations of individual u_i, but the correlation $LF\text{-}IUF(u_i, l_j)$ is no larger than sensitivity threshold ∂. In this case, the location l_j can also be released directly without violating the k-correlation demands.

(c) Location l_j is one parking location of the "top m" locations of individual u_i, and their correlation $LF\text{-}IUF(u_i, l_j)$ is larger than the sensitivity threshold ∂. This implies that l_j is much related with u_i. If one trajectory contains l_j, it is more likely that this trajectory belongs to u_i. In this case, we first take a partition of the spatial space to generate a grid G. For example, two cells c_1, c_2 are adjacent if they have a common border. IF l_j is enclosed by c_1, the *TRAMP* Algorithm merges cells to generate a new partition in which cells c_1 and c_2 are replaced by $c_1 \cup c_2$. It can be inferred that when c_1 and c_2 are merged, the correlation of the resulting cell is lower than that of c_1, that is:

$$\begin{aligned} &LF-IUF(u_i, c_1) \\ &= p_{ij} \log \frac{|U|}{\left|\left\{u_i | l_j \in c_1, u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ &\leq p_{ij} \log \\ &\quad \times \frac{|U|}{|\{u_i | l_j \in c_1 \cup c_2, u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\}|} \\ &= LF-IUF(u_i, c_1 \cup c_2) \end{aligned}$$

In essence, the coarser the partition is, the lower the correlation value between the partition and the specified user is. From this consideration, it is trivial to show that there must be a minimal merged region r satisfies the demand of k-correlation for each user:

$$\begin{aligned} &LF-IUF(u_i, r) \\ &= p_{ij} \log \frac{|U|}{\left|\left\{u_i | l_j \in r, u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ &\leq \frac{1}{m} \log \frac{|U|}{\left|\left\{u_i | l_j \in c_1, u_i \in U, tr \in tr(u_i), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \end{aligned}$$

In the worst case, the merged process degenerates returning a unique obfuscated location for the whole space. □

Theorem 2: Given a k-correlation version D^* of a trajectory dataset D and a threshold ∂, we have that, for any moving preference attacks, the probability of re-identification can be kept under ∂ by choosing the parameters k and m properly.

Proof: We use re-identification probability to measure the probability of a privacy breach of anonymized trajectory data. It represents the maximum probability of re-identification of a trajectory in anonymized trajectory data by an adversary with some background knowledge. Let $tr_i^* \in D^*$ be the anonymized version of a trajectory $tr_i \in D$. The re-identification probability of tr_i, given that a adversary's background knowledge learned from history trajectories, is calculated as:

$$P_r(u_i | tr_i^*, l_j, p_{ij}) = \begin{cases} \dfrac{1}{\left|\left\{u_q | u_q \in U, tr \in tr(u_q), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ \qquad pp \in_C l_j, pp \leftarrow tr_i^* \\ \dfrac{1}{\left|\left\{u_q | l_j \in r, u_q \in U, tr \in tr(u_q), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ \qquad pp \in_C r, pp \leftarrow tr_i^*, l_j \in r \\ 0 \quad otherwise, \end{cases}$$

where p_{ij} denotes the moving preference of user u_i, which is the probability of u_i parking on l_j gained by adversaries. Since all parking locations in tr_i^* satisfy the k-correlation, we have:

$$\begin{cases} p_{ij} \log \dfrac{n}{\left|\left\{u_q | u_q \in U, tr \in tr(u_q), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ \quad \leq \dfrac{1}{m} \log \dfrac{n}{k} \quad pp \in_C l_j, \ pp \leftarrow tr_i^* \\ p_{ij} \log \dfrac{n}{\left|\left\{u_q | l_j \in r, u_q \in U, tr \in tr(u_q), pp \leftarrow tr, pp \in_C l_j\right\}\right|} \\ \quad \leq \dfrac{1}{m} \log \dfrac{n}{k} \quad pp \in_C r, pp \leftarrow tr_i^*, \ l_j \in r. \end{cases}$$

Then the re-identification probability is:

$$P_r(u_i | tr_i^*, l_j, p_{ij}) \leq \frac{\sqrt[mp_{ij}]{n/k}}{n}.$$

Clearly, if we adjust the parameters k and m properly, we can get different privacy protection levels. □

In anonymization experiments, we randomly pick 100 taxis' trajectories to be anonymized, and the remaining trajectories are used as a history of trajectories. Figure 6 presents the total number of top m parking locations of the 100 taxis; we can then observe that a larger m leads to more correlated locations. The inherent characteristics of taxis decide that there are more parking locations than common vehicles. Generally speaking, the top 2 locations parked by common vehicles (i.e., personal cars) are usually located at home and work locations. The average number of parking locations per taxi per day is about 6.8. We, therefore, consider that the "top m" parking locations ranges from 2 to 10. When m increases, the number of parking locations we should consider increases. More than 86% locations are considered

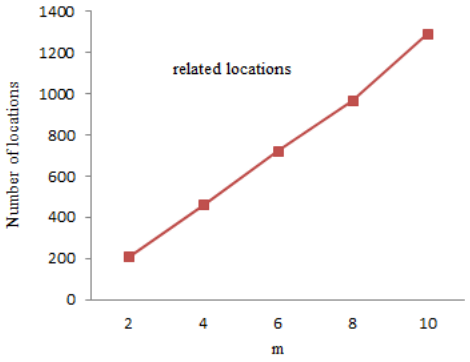


FIGURE 6. Personal-correlated top *m* locations.

in the anonymization process when m equals 10, which is adequate in our evaluation experiments.

To illustrate the privacy and impact of m and k, we choose m and k in $\{2, 4, 6, 8, 10\}$. Since the number of frequency of parking locations for most taxis is less than 10 in our dataset, it is reasonable to anonymize the trajectories while fixing m to 2, 4, 6, 8, and 10. There are 1496 locations correlated to our randomly selected 100 taxis (users). Figure 7 shows the number of the original locations published under different k.

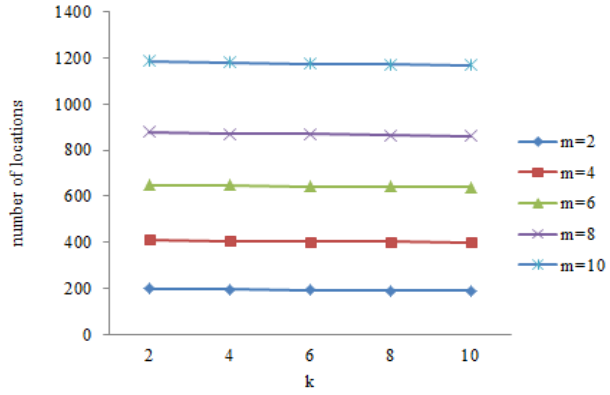


FIGURE 7. Original location published.

We can observe that there are only 209 locations related to our selected users when $m = 2$ as shown in Fig. 6, and 199 original locations are published when $k = 2$ as shown in Figure 7. In other words, only 10 locations are generalized under $k = 2$, as Figure 8 shows. When $m = 10$, there are 1291 correlated locations as shown in Fig. 6. 103 locations should be generated under $k = 2$, 110 locations were generated under $k = 4$, and 123 locations were generated under $k = 10$ as shown in Fig. 8.

Figure 8 shows that an increase of k leads to the number of generalized location increasing slowly under the fixed m. We also observe that the increase of m leads to more locations being generalized. From another point of view, m denotes the attack capability of adversaries, so as adversaries' attack capability improves, more locations needs to be generalized; and k denotes the degree of blend in the crowd, then with

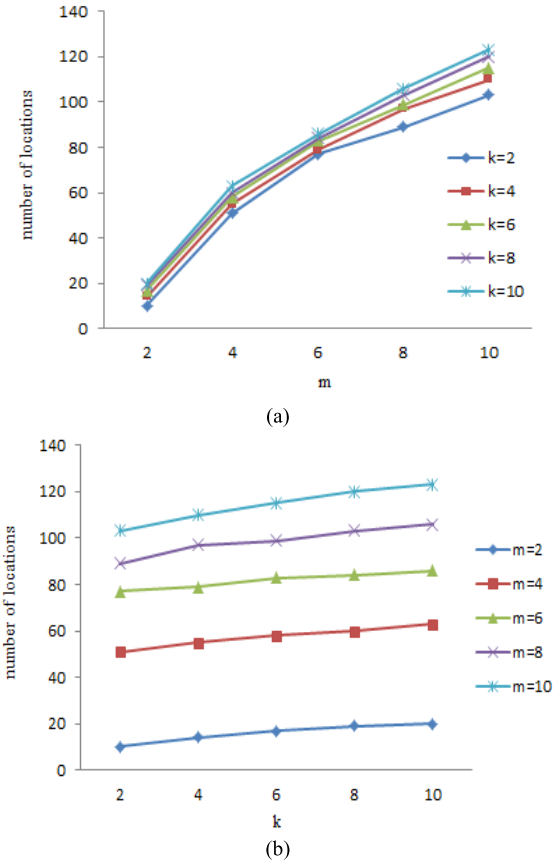


FIGURE 8. Generalized locations published. (a) numbers under differet *m*; (b) numbers under different *k*.

increasing the blend degree, also needs to generalize more locations. The main factor influencing the number of generalized locations is m, which determines privacy levels, and the other parameter k plays a fine-tuning role in the anonymization process.

For the generalized locations, we also measure their average sizes. In Figure 9, we present the average size per generalized locations under varying k from 2 to 10 while fixing m to be 2, 4, 6, 8, and 10 respectively. As expected, the average size of generalized locations increases when the privacy parameter k increases. Nevertheless, the size does not necessarily increase with the increase of m under different k. From Fig. 9, it is surprising to see that the average size with $m = 6$ and $k = 6$ is 48.99, while the generalized average sizes are 47.33 under $m = 8, k = 6$, and 44.17 under $m = 10$, $k = 6$ respectively. This is because the number of generalized locations with $m = 6$ is less than those of the cases $m = 8$ and $m = 10$. When calculating the average size, fewer locations share the sum of generalized locations size with $m = 6$, which can easily make the average size of the generalized location relatively larger.

Next we study how information loss vary under different correlation parameters while fixing $m = 2$, $m = 4$, $m = 6$, $m = 8$, $m = 10$ and varied k in [2] and [10].

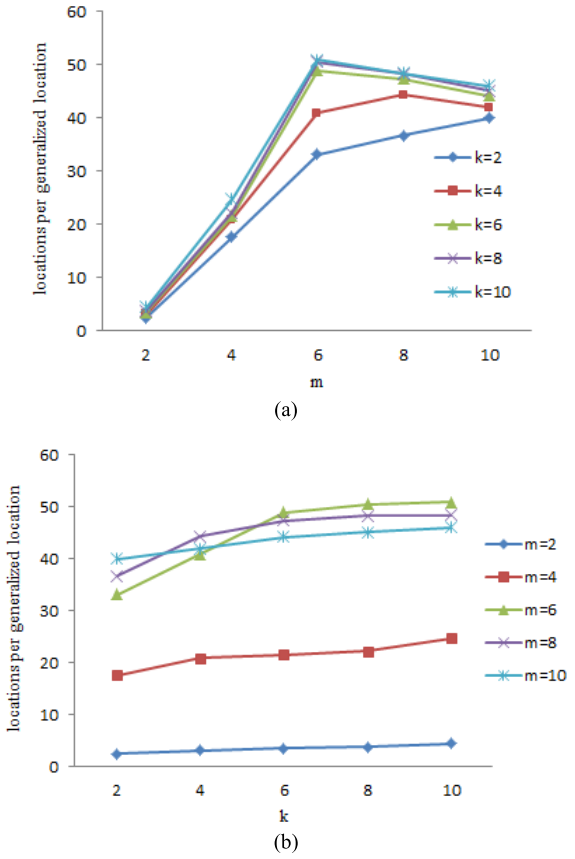


FIGURE 9. Top 10 parking locations Average size of generalized locations. (a) size under different *m*; (b) size under different *k*.

C. MEASURE OF INFORMATION LOSS

To quantify information loss, we measure the number of original and generalized locations in an anonymized dataset. For the generalized locations, we measure the average information loss by (6), similar to [2] and [9].

Figure 10 examines the information loss for different correlation parameters k and m under three different anonymization methods, k-anonymity, our method *TRAMP*, *GridPartition* proposed in [2].

In general, the information loss increases by the increment of k. Comparison of all the three algorithms are shown in Fig.10, we can see that the *GridPartition* causes more information loss than that of k-anonymity and *TRAMP*, but the information loss caused by *TRAMP* is slightly higher than that of k-anonymity. Since k-anonymity adopts a clustering strategy, which completely ignores the correlation of locations, it only needs to meet that at least k different moving objects co-exist in a location or a generalized location, and this condition is easy to satisfy when the volume of data is large, but it cannot defend against moving preference attack.

The information loss caused by *GridPartition* is mainly caused by generalization of parking points, which generalizes all parking locations in its anonymization process, however *TRAMP* states that not all parking locations should be satisfy

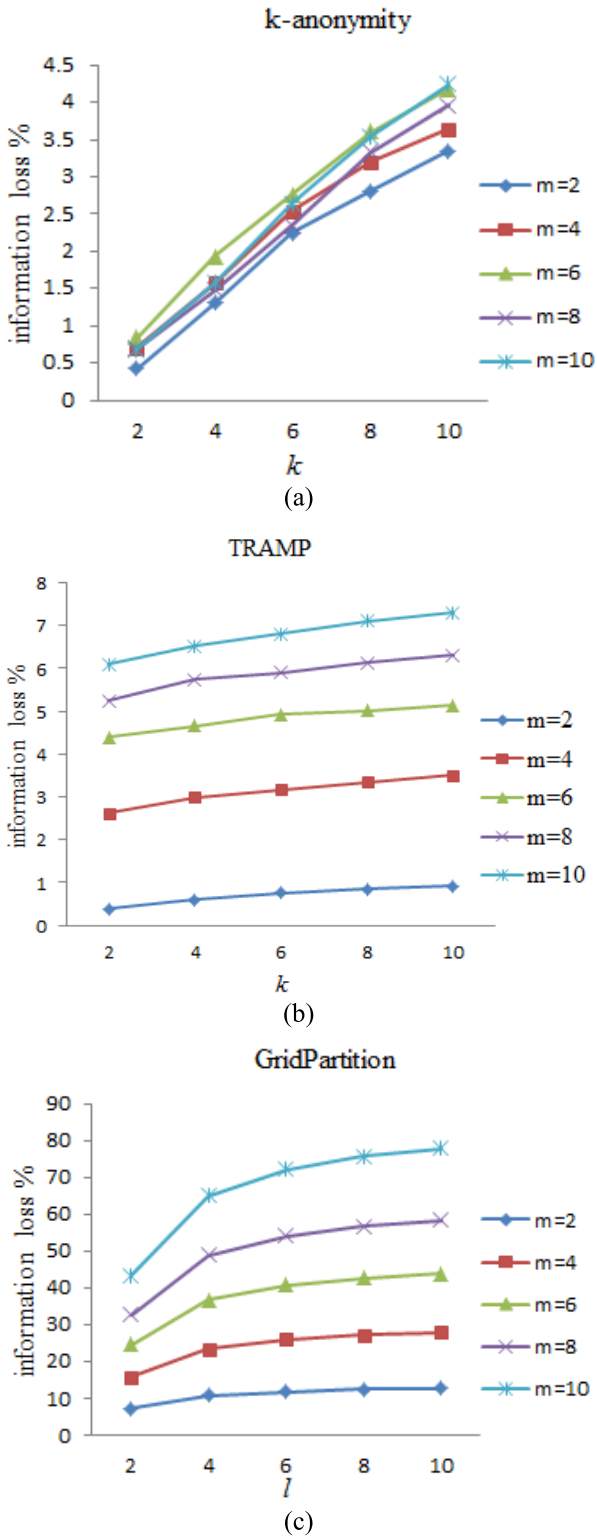


FIGURE 10. Information loss. (a) loss of *k*-anonymity; (b) loss of *TRAMP*; (c) loss of *GridPartition*.

the demands of k-anonymity or l-diversity; in other words, it generalizes only these parking locations with high correlation to individuals. Therefore, *TRAMP* performs better

than *GridPartition* on Information loss. In all three figures, performance decreases as privacy level grows. However, the increment of k in k-anonymity and increment of l in *GridPartition* will make a great change in the information loss as shown in Fig.10 (a) and (c), while increment of k in *TRAMP* will only make a very small change in the information loss as plotted in Fig.10 (b).

Therefore, in our proposed method, *TRAMP, k* slightly affects the information loss. In other words, less information loss is added when improving the privacy level through increasing k. In addition, the information loss in *TRAMP* increases with an increase of m, because more correlated parking locations need to be generalized to hide stronger correlation, which increase information loss. We can observe that a large m can provide high privacy level even when k is relatively small in *TRAMP*. Meanwhile it causes more information loss. In practice, we can select a proper parameter pair of (k, m) to balance the privacy and information loss. The experiment's results not only illustrate the validity of our proposed privacy framework, but also provide guidance on how to set privacy parameters m and k.

VI. CONCLUSION

With development of information technology, the IoT and big data have integrated into various aspects of human life. As such, security and privacy are becoming more difficult to handle as the IoT is characterized by numerous mobile devices and vulnerability to many security attacks. Many works in the IoT discuss the privacy and security of communication channels, user authentication and authorization, but few works focus on the privacy of data itself. The huge amount of collected data contains profoundly sensitive information. This paper contributes in the field of individual privacy of trajectory data generated in intelligent transportation systems, which is a main application of IoT. We analysed location trajectory from the perspective of individual's privacy.

A generalized model to balance trajectory privacy and information loss was studied. Most trajectory methods ignore individual's distinguishability in terms of moving preference. However, adversaries may use the uniqueness of a moving object to identify anonymized trajectories. Moreover, the existing methods, such as [2], generalize all parking locations no matter how they are related to an individual; this leads to an increase of information loss along with the expansion of all parking locations. To overcome these limitations, we took an individual's history trajectories into account, and learned an individual's moving preference from them. We then proposed to use personal-correlated location to represent the moving preference, and developed a new anonymized framework that takes the correlation between locations and individuals into account. For each trajectory to be published, we calculated every single location's correlation using *LF-IUF*, which is related to the studied individual and other individuals who have been parked at that location. Consequently, for each personal-correlated location, we determined a k-correlation region to replace it with the anonymized trajectories. Our approach created a new method to evaluate the correlations between locations and individuals. It is a practical solution for trajectory data publication via generalization. Extensive experiments demonstrated that our solution performs well in terms of privacy protection and information loss. The decision of generalizing a correlation parking location is done through the *LF-IUF* value differences, that is, if the correlation is very small, then it is not necessary to generalize, thus avoiding over-protection problems. In addition, our method can effectively resist individual's moving preference attacks.

However, in the context of trajectory data the privacy protection is very challenging, because location is a special kind of privacy that contains abundant spatiotemporal information. As future work, we are planning to apply more sophisticated method such as real-time streaming data analysis, trajectory compression, road network matching, frequent sub-trajectory mining, and other spatiotemporal processing techniques to find effective solutions for privacy protection of moving object data.

ACKNOWLEDGMENT

Guangxi Bagui Scholar Teams for Innovation and Research Project and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

REFERENCES

[1] P. P. Jayaraman, X. Yang, A. Yavari, D. Georgakopoulos, and X. Yi, "Privacy preserving Internet of Things: From privacy techniques to a blueprint architecture and efficient implementation," *Future Generat. Comput. Syst.*, vol. 76, pp. 540–549, Nov. 2017.

[2] Z. Huo, X. Meng, H. Hu, and Y. Huang, "You can walk alone: Trajectory privacy-preserving through significant stays protection," in *Proc. 17th Int. Conf. Database Syst. Adv. Appl.*, Busan, South Korea, 2012, pp. 351–366.

[3] Y. Zheng and X. Zhou, *Computing With Spatial Trajectories*. New York, NY, USA: Springer, 2011.

[4] M. E. Nergiz, M. Atzori, Y. Saygın, and B. Güç, "Towards trajectory anonymization: A generalization-based approach," *Trans. Data Privacy*, vol. 2, no. 1, pp. 47–75, 2009.

[5] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition using relational Markov networks," in *Proc. 19th Int. Joint Conf. Artif. Intell.*, Edinburgh, Scotland, 2005, pp. 773–778.

[6] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *IEEE Pervasive Comput.*, vol. 5, no. 4, pp. 38–46, Oct. 2006.

[7] J. Krumm, "Inference attacks on location tracks," in *Pervasive Computing*. Berlin, Germany: Springer, 2007, pp. 127–143.

[8] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. 24th Int. Conf. Data Eng.*, 2008, pp. 376–385.

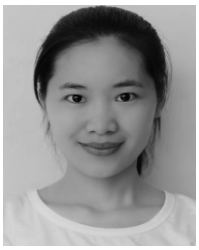
[9] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: How to hide a MOB in a crowd?" in *Proc. 12th Int. Conf. Extending Database Technol.*, Saint Petersburg, Russia, 2009, pp. 72–83.

[10] J. Domingo-Ferrer and R. Trujillo-Rasua, "Microaggregation- and permutation-based anonymization of movement data," *Inf. Sci.*, vol. 208, no. 21, pp. 55–80, 2012.

[11] N. Mohammed, B. C. M. Fung, and M. Debbabi, "Walking in the crowd: Anonymizing trajectory data for pattern analysis," in *Proc. 18th Int. Conf. Inf. Knowl. Manage.*, 2009, pp. 1441–1444.

[12] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, "Apriori-based algorithms for k^m-anonymizing trajectory data," *Trans. Data Privacy*, vol. 7, no. 2, pp. 165–194, 2014.

[13] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proc. 9th Int. Conf. Mobile Data Manage.*, 2008, pp. 65–72.
[14] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Inf. Sci.*, vol. 231, pp. 83–97, May 2013.
[15] R. Chen, B. C. M. Fung, and B. C. Desai. (2011). "Differentially private trajectory data publication." [Online]. Available: https://arxiv.org/abs/1112.2020
[16] R. Chen, Q. Xiao, Y. Zhang, and J. Xu, "Differentially private high-dimensional data publication via sampling-based inference," in *Proc. 21st Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 129–138.
[17] A. Monreale, R. Trasarti, D. Pedreschi, C. Renso, and V. Bogorny, "C-safety: A framework for the anonymization of semantic trajectories," *Trans. Data Privacy*, vol. 4, no. 2, pp. 73–101, 2011.
[18] E. Yigitoglu, M. L. Damiani, O. Abul, and C. Silvestri, "Privacy-preserving sharing of sensitive semantic locations under road-network constraints," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage.*, Jul. 2012, pp. 186–195.
[19] P. S. Castro, D. Zhang, and S. Li, "Urban traffic modelling and prediction using large scale taxi GPS traces," in *Pervasive Computing*, vol. 7319. Berlin, Germany: Springer, 2012, pp. 57–72.
[20] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 126–146, 2008.
[21] P. Sui, T. Wo, Z. Wen, and X. Li, "Privacy-preserving trajectory publication against parking point attacks," in *Proc. IEEE 10th Int. Conf. Ubiquitous Intell. Comput.*, Dec. 2013, pp. 569–574.
[22] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 791–800.
[23] A. E. Cicek, M. E. Nergiz, and Y. Saygın, "Ensuring location diversity in privacy-preserving spatio-temporal data publishing," *Int. J. Very Large Data Bases*, vol. 23, no. 4, pp. 609–625, 2014.
[24] Y. Song, D. Dahlmeier, and S. Bressan, "Not so unique in the crowd: A simple and effective algorithm for anonymizing location data," in *Proc. 1st Int. Workshop Privacy-Preserving Inf. Retr.*, 2014, pp. 19–24.
[25] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. 18th Int. Conf. Mobile Comput. Netw.*, 2011, pp. 145–156.
[26] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, 2014.
[27] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao, "Privacy vulnerability of published anonymous mobility traces," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 720–733, Jun. 2013.
[28] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. 32nd Int. Symp. Secur. Privacy*, 2011, pp. 247–262.
[29] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 6, p. 1376, Mar. 2013.
[30] J. Hua, Z. Shen, and S. Zhong, "We Can Track you if you take the metro: Tracking metro riders using accelerometers on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 2, pp. 286–297, Feb. 2017.



PEIPEI SUI was born in 1987. She received the B.S. and M.S. degrees from Yanshan University, China, in 2007 and 2010, respectively, and the Ph.D. degree in computer science from Beihang University, China, in 2017. She is currently a Lecturer with the School of Management Science and Engineering, Shandong Normal Univeristy, China. Her current research interests include data analysis, spatial-temporal data mining, and data privacy.

XIANXIAN LI was born in 1969. He received the Ph.D. degree in computer science from Beihang University, China, in 2006. He is currently the Dean and a Professor with the Department of Information Science and Engineering, Guangxi Normal University. His research interests mainly include network and multi-source information security, trusted computing, and data privacy.

YAN BAI received the Ph.D. degree in electrical and computer engineering from The University of British Columbia, Vancouver, BC, Canada. She is currently an Associate Professor with the University of Washington Tacoma. Her research interests are in the areas of cyber security, computer networking, eHealth and health IT, cloud computing, and multimedia communications.