Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 5850670, 13 pages
http://dx.doi.org/10.1155/2016/5850670

Research Article

On Preventing Location Attacks for Urban Vehicular Networks

Meng Zhou,[1,2] Xin Li,[1,2] and Lejian Liao[1]

[1]*School of Computer Science, Beijing Institute of Technology, Beijing 100081, China*
[2]*Beijing Engineering Application Research Center of High Volume Language Information Processing and Cloud Computing, Beijing 100081, China*

Correspondence should be addressed to Xin Li; xinli@bit.edu.cn

Received 25 June 2016; Revised 22 September 2016; Accepted 23 October 2016

Academic Editor: Hui Zhu

The prevalence of global positioning system (GPS) equipped in vehicular networks exposes users' location information to the location-based services. We argue that such data contains rich informative cues on drivers' private behaviors and preferences, which will lead to the location privacy attacks. In this paper, we proposed a sophisticated prediction model to predict driver's next location by using a k-order Markov chain-based third-rank tensor representing the partially observed transfer information of vehicles. Then Bayesian Personalized Ranking (BPR) is used to learn the unobserved transitions within the tensor for transition predication. Experimental results manifest the efficacy of the proposed model in terms of location predication accuracy, compared with several state-of-the-art predication methods. We also point out that the precision achieved by such advanced predication model is restricted to the order of the Markov chain k. Accordingly, we propose a schema to decrease the risks of such attacks by preventing the conformation of higher order Markov chain. Experimental results obtained based on the real-world vehicular network data demonstrated the effectiveness of our proposed schema.

1. Introduction

With the prevalence of global positioning system (GPS) and vehicular networks, the usage of smart phones and in-car navigation systems plays an increasingly important role in our daily lives. While enjoying the convenience brought by various location-based services (LBSs), such as mapping, route finding, and automotive traffic monitoring, people inevitably release their physical location information for public access. Unfortunately, such disclosure of location information consequently induces serious privacy issues [1–3]. For example, some social networks users need to report their sequential locations to a service provider in a periodic or on-demand manner to obtain its desired location-based services, for example, advertising and restaurant recommendation, while the disclosed personal location data may be used for location privacy attacks by adversaries. Thus, location privacy protection in LBSs has attracted a lot of research attention from both industry and academia. The location privacy concern does appear not only in mobile social networks, but also in vehicular networks. Essentially, vehicular networks are similar to mobile social networks in terms of mobility, connectivity, and ubiquity. Vehicles are equipped with wireless sensor devices and modeled as moving nodes, forming networks for vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) [4]. Signals captured from these moving nodes can then be used to detect road conditions such as traffic flow and traffic signals [5]. The vehicular networks are expected to play critical roles in our daily life such as road infrastructure monitoring, driving risk detection, and passenger communication.

However, adversaries can also use the location information to estimate users' private information such as social ties [6] or personal activities [7]. It is more dangerous if adversaries use such inference to carry out *physical* location attacks. Privacy preservation is an important component of customer-centric pervasive services. Without the guarantee of privacy protection, users would be hesitant to use LBSs which continuously monitor their locations [3]. Several research studies have been proposed to protect location privacy from being disclosed through inference-based approaches, such as *K-anonymity* [8], *pseudonyms and*

mix-zones [9], and *path confusion* [10]. These approaches anonymize accurate information of users and make them indistinguishable among the neighboring users. While strengthening users' privacy, these methods in turn weaken the functionality of the service by updating the inaccurate spatial information or information with the high latency on purpose [11].

Another potential privacy threat is that the current location information of users can be inferred from the historical data even though they are not directly disclosed. Wu et al. have demonstrated that there is a strong spatiotemporal regularity with vehicle mobility through a conditional entropy analysis [12], which indicates that the ability of prediction of future locations must be considered when we cope with the location privacy attacks. We can imagine that the adversary could ambush a vehicle by using the location prediction method to infer the possible future locations. In this paper, we will focus on preventing those potential attacks from a perspective of location prediction. A common scenario is that the adversary can predict user's next location at time t using location prediction models for privacy attacks. Our objective is then to propose a schema to prevent the vehicles from such privacy attacks.

In the literature, there exist a lot of works related to location predication. In [13], the authors proposed to use GPS traces to infer the mode of a personal transportation and then to predict their routes based on people's historical trajectory data. Other works include determining which road a driver is on in spite of the noisy GPS data [14] and predicting the destination of a trip [15]. In [16], the author also pointed out a lot of personal information can be inferred from their long-term location history, for example, age, work role, work group, work frequency, coffee drinker, smoker, work room, and which train station they favored. However there are few works addressing the location predication problem from a perspective of privacy protection. In fact, the report from US Department of Justice had revealed that approximately 26,000 persons are victims of GPS stalking annually [17]. All the aforementioned could provide the opportunities for adversaries to launch a location-based attack, as shown in Figure 1. Assuming that adversaries can predict the target vehicle's future location based on the effective prediction approach, the adversaries then can be well prepared near the predicted location to ambush the coming target.

Under the aforementioned scenarios, we argue that the location privacy problem in vehicular networks is a prediction related problem. Some research on location privacy attacks is proposed on top of a Markov chain (MC) model [18, 19], which has demonstrated their effectiveness on the prediction results. In this paper, we proposed a more sophisticated predication model, that is, a k-order Markov chain-based tensor model, to predict the future locations of vehicles. The successive locations can be obtained by using a Bayesian Personalized Ranking (BPR) approach. Our experimental results show that the predication accuracy achieved by our proposed approach is much higher than that of the conventional approach. This indicates that a carefully designed learning approach can successfully discover the partially observed transitions between a number of successive

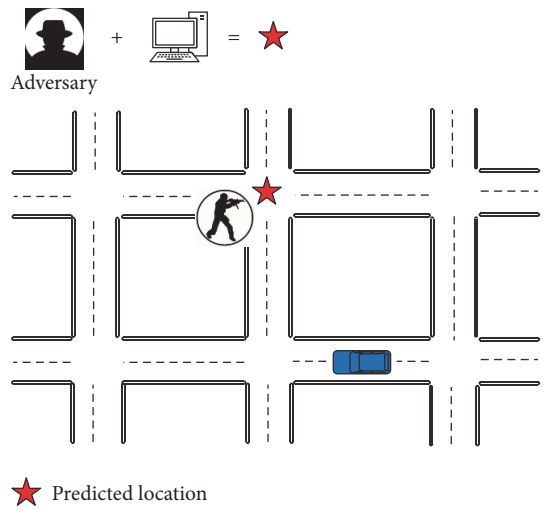


Figure 1: An example of location attack.

next locations and the missing transitions. Then, we analyze how the order of k can affect the prediction accuracy of MC-based model. At last, we propose a schema by setting up a reasonable time slot to prevent vehicles from conforming the higher order Markov chain to lower the predication accuracy; thus the vehicle is protected from the risk of location attacks.

The key contributions of this work can be summarized as follows:

(i) We proposed a sophisticated location prediction model which will be referred to as k-order FPMC in this paper.

(ii) We analyzed the key factors of the proposed prediction model to the prediction accuracy and pointed out that the location privacy problem in vehicular networks is restricted to the order k of Markov chain and proposed a strategy to protect the privacy leaking from such predication model.

(iii) We evaluated the proposed approach using the real-world traffic trace data. Experimental results obtained manifest the efficacy of our proposed approach.

The rest of the paper is organized as follows. Section 2 reviews the related works on location privacy and location prediction. The system model and the adversarial model adopted in this paper are introduced in Section 3. We propose a novel approach for mining vehicles' trajectory as well as the strategies against privacy attack in Sections 4 and 5, respectively. Section 6 presents the experimental results and we conclude the paper in Section 7.

2. Related Works

In this section, we briefly review the existing literatures with a focus on recent developments in location prediction and location privacy protection, respectively.

For location prediction, Krumm [20] proposed to predict drivers' turn proportions at road intersections at a fine-grained level. The idea is to choose the higher likelihood

on a turn that links more destinations. Veloso et al. [21] proposed to utilize a Naive Bayes model to predict the relation between pick-up and drop-off locations, and their work also explored the possibility to predict area type of the next pick-up location, given the features of drop-off location, for example, time and day, weather condition, and area type of current drop-off location. Ziebart et al. [22] proposed to model observed behaviors by learning context-aware action utilities for turn prediction, route prediction, and destination prediction. Wu et al. [12] proposed to develop an efficient data delivery by predicting vehicle trajectories via multiple order Markov chain. Qin et al. [23] studied the mobile advertising problem in vehicular network and proposed to adopt Markov chain to capture the patterned vehicular centrality and to infer the future traffic flow. Chen et al. [24] analyzed the predictability of taxi mobility via a Markov predictor.

All the aforementioned work built up their predication model on top of Markov chain to facilitate the trajectory prediction, which failed to solve prediction problem with cold start issue, as the conventional Markov chain model cannot work on things never happened before. In our model, we take the preference between locations into consideration; thus we can acquire an average preference from other users between two locations that the user never leaves a footprint.

For the issue of location privacy, the related attacks have been studied in the literature. Location prediction attacks proposed by Minami and Borisov [18] studied an issue of inference attacks on the GPS traces. The analysis revealed that if there is an adversary who could access to a mobile user's previous location data, a Markov model-based location predictor could be adopted to assist the next location attack.

Shokri et al. [19] formalized a sporadic location-based application and found that an adversary who knows personalized transition matrices of the users could deanonymize and deobfuscate traces with higher accuracy than an adversary who only knows each user's prior probabilities on locations. De Mulder et al. [25] demonstrated that it is possible to build up the profiles of users' movements based on the GSM location data, which lead to identify the users in a subsequent period with great accuracy (about 80% of the time). The location profile model used is a simple first-order Markov chain. Gambs et al. [26] designed the novel distances to quantify the similarity between two MMCs and described how these metrics can be combined to build deanonymizers. These three methods also used Markov chain to carry out the prediction, also met the problem we mentioned before, and will be compared with our proposed approach at the section of experiments.

The deanonymization attack is very accurate with a success rate of up to 45% on the Geolife dataset [27]. It reminds us that the anonymizers can be deanonymized by a specific method; therefore it is necessary to lower the attack success rate after the adversary knows who you are. While our proposed method allows the adversary to know who you are, it is hard for the adversary to determine where you are going.

There are also some works that focused on the defenses by providing only partial data about the users' locations and identities [11]. K-anonymity [8] provides a form of plausible deniability by ensuring that the user cannot be individually identified from a group of k users. This can be achieved in vehicular networks by setting a large k-anonymous region which included k users, instead of just reporting a single GPS location.

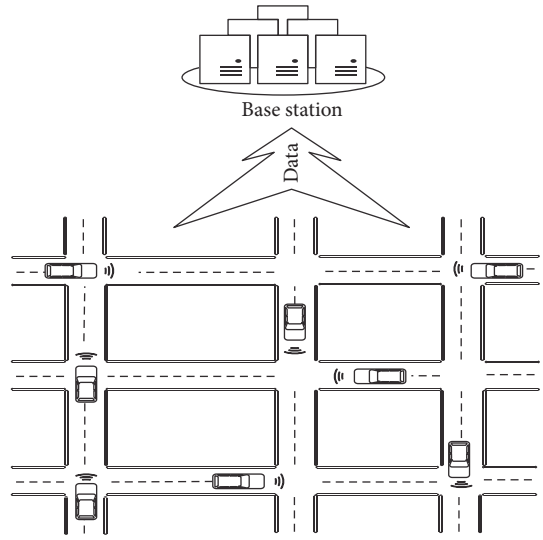


Figure 2: System model.

Pseudonyms and mix-zones [9] provide a certain degree of anonymity to the individual user. When the users enter a mix-zone, they change a new, unused pseudonym. In addition, they do not send their location information to any location-based application when they are in the mix-zone. Mix-zones also impose the limits on the frequent updating, that is, the exposure of the pattern of closely spaced queries, allowing one to easily follow the user. Path confusion [10] avoids linking consecutive location samples to individual vehicles through target tracking algorithms with high certainty. The main impediment to the use of path confusion is the processing delay; one must wait until users' paths have intersected before revealing those locations to a location-based service.

In this paper, we propose a sophisticated model to reveal that the location privacy could be obtained by the prediction model with high success rate. The experimental results and the analysis over the real vehicle GPS traces data collected in a mega city, Shanghai, China, suggest that we need to use an obfuscation approach instead of anonymization method, by hiding previous location to avoid releasing consecutive information to protect the user's location privacy from such advanced prediction model.

3. Models and Goals

We define the system model, the adversarial model, and the security goal in this section. Our system model consists of two components, that is, vehicles and base station, which is shown in Figure 2. Basically, the vehicles distribute their trajectory information to the base station constantly.

3.1. Adversarial Model. We also consider two kinds of adversarial models in the proposed system. One is outside adversary model, and another one is inside adversary. From

the point of outside attacker, he can listen, insert, delete, and modify the communication message between the base station and vehicles in the system. This threat can be avoided by adopting conventional entity identity authentication and key exchange protocols [28], such as SSL [29] and TLS [30]. More serious threats are posed from inside attackers. For instance, a database administrator in the base station might sell customers' historical location information to certain data analytics companies, in order to have financial income. Since this attacker could access customers' sensitive information directly, the conventional privacy preservation approaches are therefore challenged a lot.

3.2. Goals. There are two goals in this system, that is, the security goal and the prediction goal.

In this paper, the prediction goal is simply measured by the prediction accuracy. And the security goal is defined as follows.

Definition 1. The vehicular location achieves (p, t) location privacy, if the successful prediction rate is less than p in time t.

In Section 4, we present the details of our proposed predication model.

4. Tensor-Based Location Prediction Framework

In this section, we present our location prediction framework in great detail. We first describe the temporal characteristics of the trajectory and the motivation of adopting tensor to represent the trace data characterizing the temporal relations of data in Section 4.1. We then propose to adopt a tensor factorization approach towards recovering the missing data in Section 4.2. Finally, we describe the learning process of the prediction model by using BPR criteria in Section 4.3.

Here we introduce notations used throughout this paper. Let $V = \{v_1, v_2, \ldots, v_{|V|}\}$ denote a set of vehicles. As taxies are randomly distributed in the city running along the roads and are constrained by road conditions, we thus denote trajectories of taxies by using the successive crossroads they have passed. Let $L_V = \{1, 2, \ldots, n\}$ denote a set of crossroads, where each cross is geocoded by {longitude, latitude}, and n is the total number of crossroads. For each taxi v, the historical trajectory is denoted as $L^v := \{L_1^v, L_2^v, \ldots, L_{t-1}^v\}$ with $L_t^v \subseteq L^v$, and t is the time slots, and $L := \{L^{v_1}, L^{v_2}, \ldots, L^{v_{|V|}}\}$ denotes the trajectories over the entire set of vehicles. And each road segment is labeled by its adjacent crosses.

4.1. High Order MC Representation via Tensor. In this section, we describe the details of the construction of third-rank tensor, each item of which represents the approximate probability of transferring from a specified combination of intersections (locations) to another intersection (location) for a particular vehicle.

An m order Markov chain is defined as

$$\Pr\left(X_t = x \mid X_{t-1} = x_{t-1},\ X_{t-2} = x_{t-2}, \ldots,\ X_{t-n} = x_{t-n}\right), \quad (1)$$

where $X_{t-1}, \ldots, X_{t-n}$ is a sequence of random variables and x_{t-n} is their realizations. Markov chain models sequential behaviors by learning a transition graph over items, and thus Markov chain can be directly adapted to predict the future locations based on one's recent trajectory data. Our proposed model is to predict the next personalized location via the ranking of probabilities that a vehicle v will move from its current location to the next location. In vehicular networks, network topology changed rapidly and our target is to investigate the main factors affecting the prediction accuracy. Unfortunately, the first-order Markov chain is unable to capture the rapid changing features in this scenario. Consequently, we extend the low order Markov chain to a high order version. Assuming that the order of Markov chain is set to 3, we will predict the next location of taxi v according to its three previous locations. The transition probability from current location to next location can be written as

$$p\left(L_{t+1} = l_{t+1} \mid L_t = l_t,\ L_{t-1} = l_{t-1},\ L_{t-2} = l_{t-2}\right), \quad (2)$$

where l_{t+1} is the next location of taxi v, l_t is the current location of v at time t, and l_{t-1}, l_{t-2} are the previous location of v at time $t-1$ and $t-2$, respectively. Let C denote the set of current locations c and $c = \{L_{t-1} = l_{t-1},\ L_{t-2} = l_{t-2},\ L_{t-3} = l_{t-3}\}$. The probability of next location can be deduced from the current location as $p(L^v|C^v)$. And we define each item of the 3-order tensor is the transition probability between two adjacent crosses denoted as $x_{v,c,l}$ and $x_{v,c,l} = p(l^v|c^v)$. Then, each vehicle is associated with a specific transition matrix χ^v, and a transition tensor is yielded from all the vehicles. More specifically, $S \subseteq V \times C \times L$ in our paper is used to represent the transitions for each taxi v, where C and L denote the set of the previous locations and next locations, respectively. $\chi_{v,c,l}$ represents the observed transitions of v from c to location l.

Given the location set L^v, $x_{v,c,l}$ can be estimated as follows:

$$\begin{aligned} x_{v,c,l} &= p\left(l \in L_v \mid c \in C_v\right) = \frac{p\left(l \in L_v \wedge c \in C_v\right)}{p\left(c \in C_v\right)} \\ &= \frac{p\left(l_{t+1} \in L_{t+1}^v \wedge l_t \in L_t^v \wedge l_{t-1} \in L_{t-1}^v \wedge l_{t-2} \in L_{t-2}^v\right)}{p\left(l_t \in L_t^v \wedge l_{t-1} \in L_{t-1}^v \wedge l_{t-2} \in L_{t-2}^v\right)} \\ &= \frac{\left|\left\{(L^v) : l_{t+1} \wedge l_t \wedge l_{t-1} \wedge l_{t-2}\right\}\right|}{\left|\left\{(L^v) : l_t \wedge l_{t-1} \wedge l_{t-2}\right\}\right|}. \end{aligned} \quad (3)$$

Figure 3(b) illustrates the location transitions for a particular vehicle. Figure 3(a) plots a snapshot of the road network which contains seven intersections and there is a taxi v passing by the location C at time t. Assuming the order of Markov chain is set to 3, we need to consider 3 previous locations in our model and construct the transition matrix between the combinations of three locations and the next location. Considering the entire set of the locations, there will be way too many such combinations. In fact, the number of the combinations of road intersections is constrained by the layout of the road network. Thus, in this paper we only take the practical combinations into consideration. The values of matrix items shown in Figure 3(b) are obtained according to the number of occurrence of the transitions from a particular

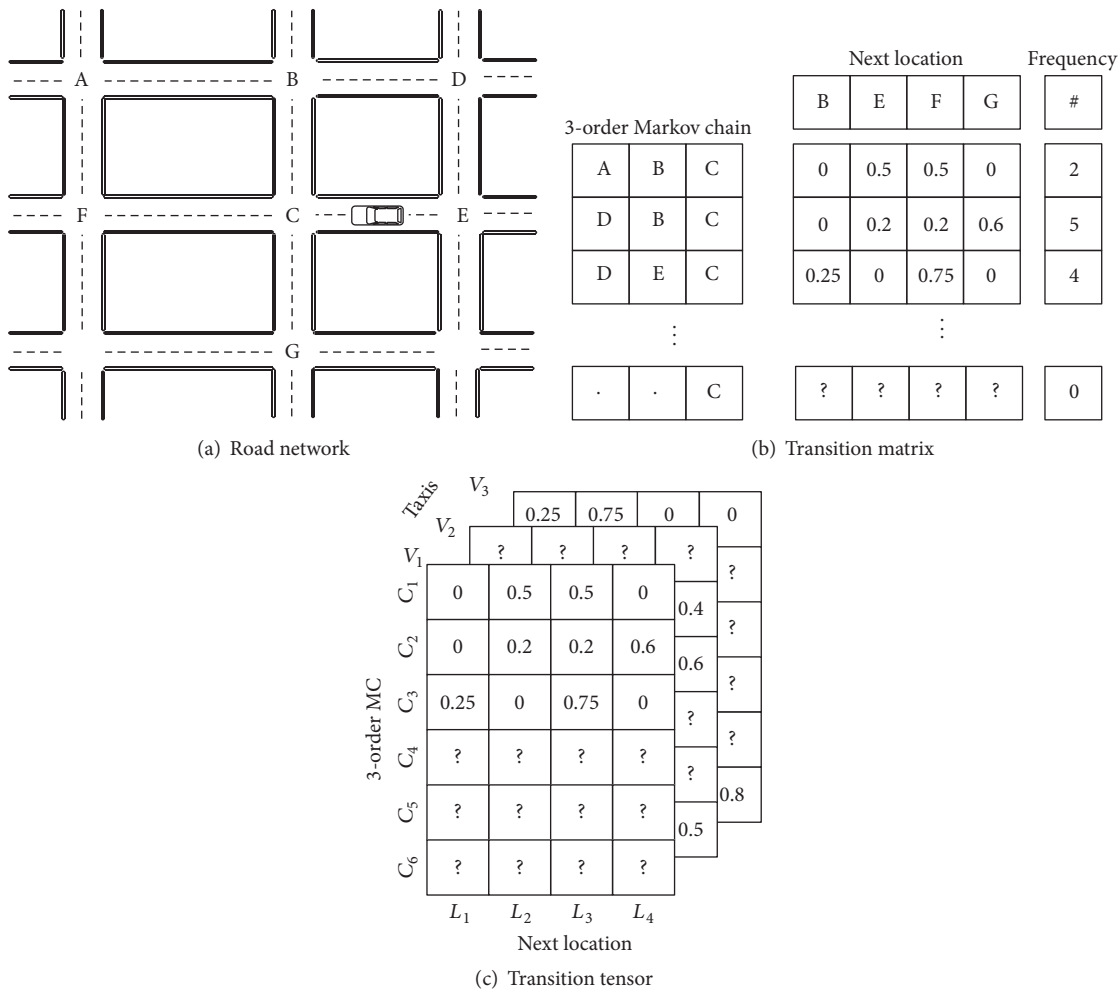


Figure 3: A Third-rank transition tensor construction.

combination of intersections to a location in the observed historical data of a vehicle. "?" denotes the missing value or the unobserved data. It is natural to extend the transition probability matrix to be a third-rank tensor by incorporating the personalization dimension as shown in Figure 3(c). Then, our objective is to infer the proper tensor model from the observed transitions to estimate the transition preference for those unobserved transition pairs.

4.2. A Pairwise Interaction-Based Tensor Factorization. In this section, we demonstrate the details of our adapted pairwise interaction-based tensor factorization towards inferring the unobserved data within the previously constructed tensor.

As aforementioned, the transition tensor is only partially observed. Similar to the conventional matrix factorization approach which infers the missing data via factorized latent features, here we adopt the low-rank factorization model to fill in missing values. The matrix factorization methods, for example, nonnegative matrix factorization (NMF) and singular value decomposition (SVD), have been successfully applied to many applications such as image processing, link prediction, and rating prediction. For the tensor factorization, tucker decomposition (TD) and canonical decomposition (CD) are the two popular techniques which have been successfully used for tag recommendation. In this paper, we adopt a special case of canonical decomposition proposed in [31] which is called the pairwise interaction tensor factorization model (PITF).

With the PITF model, we can easily model the pairwise interactions among all three components of the third-rank tensor (i.e., vehicle V, Markov chain C, and next location L), written as

$$\widehat{x}_{v,c,l} = \sum_{f} \widehat{v}^{L}_{v,f} \cdot \widehat{l}^{V}_{l,f} + \sum_{f} \widehat{c}^{L}_{c,f} \cdot \widehat{l}^{C}_{l,f} + \sum_{f} \widehat{v}^{C}_{v,f} \cdot \widehat{c}^{V}_{c,f}, \quad (4)$$

where $\widehat{v}^{L}_{v,f}$ and $\widehat{l}^{V}_{l,f}$ represent the fth latent feature for vehicle v and next location l, respectively. In (4), a tensor χ is factorized into three pairwise interaction matrix models. Note that the pairwise interaction between vehicle and its 3-order history trajectory is not related to the prediction of next location, and thus we remove the term of $\widehat{v}^{C}_{v,f} \cdot \widehat{c}^{V}_{c,f}$ as it is independent of l

as shown in [32]. Accordingly, (4) can be rewritten in a more compact form as follows:

$$\widehat{x}_{v,c,l} = \sum_{f} \widehat{v}^{L}_{v,f} \cdot \widehat{l}^{V}_{l,f} + \sum_{f} \widehat{c}^{L}_{c,f} \cdot \widehat{l}^{C}_{l,f}. \tag{5}$$

The parameter set is given as

$$\begin{aligned} \widehat{V} &\subseteq \mathbb{R}^{|V|\times F}, \\ \widehat{C} &\subseteq \mathbb{R}^{|C|\times F}, \\ \widehat{L}^{V} &\subseteq \mathbb{R}^{|L|\times F}, \\ \widehat{L}^{C} &\subseteq \mathbb{R}^{|L|\times F}, \end{aligned} \tag{6}$$

where F is the dimension of the latent feature space. And hereafter we use Θ to represent the parameter set $\{V, C, L^V, L^C\}$.

4.3. BPR Learning Criterion. In this section, we detail the learning process of the unobserved data of tensor by using BPR criteria and the stochastic gradient descent algorithm collaboratively.

The objective of location prediction is in fact to select the most likely arrived location among all the locations. An alternative way is to derive a proper ranking $>_{v,t}$ of the possibility over the candidate locations. We adopt the sequential Bayesian Personalized Ranking (S-BPR) optimization criterion [33] here. S-BPR regards the rating prediction problem as a ranking problem and assumes every two locations have a sequential relation $l_i >_{v,t} l_j$, written as

$$l_i >_{v,t} l_j : \Longleftrightarrow \widehat{x}_{v,t,l_i} > \widehat{x}_{v,t,l_j}. \tag{7}$$

Equation (7) indicates that given the current Markov chain c, if v has visited location l_i more frequently than location l_j, the probability that v visits l_i via c is bigger than that of visiting l_j. However, the conventional implicit feedback-based learning approach BPR can only infer that vehicle v prefers l_i compared to l_j without knowing the scale of the preference. For example, although we know that v prefers l_i to l_z, we cannot tell how much is the difference between l_j and l_z. We can only drive the preference pair $l_i >_{v,t} l_j$, $l_i >_{v,t} l_z$. To address such issue, we incorporate BPR with confidence as proposed in [34, 35] to model the difference and return each feedback with a confidence weight. The frequency of traveling between locations was adopted to build the preference pairs. For instance, given c, the vehicle v has been to the location l_i twice and turns to l_j once. We could conclude that the vehicle prefers to go to l_i rather than to l_j. Consequently, if we know this vehicle has never turned to location l_z, we are intuitively more confident to generate the pair of $l_i >_{v,t} l_z$.

We propose a confidence score $C_{\langle v,i,j\rangle}$ to measure to what extent v prefers location l_i to location l_j which is given as

$$C_{\langle v,i,j\rangle} = 1 - \frac{T_i - T_j}{T_i + T_j}, \tag{8}$$

where T_i and T_j are the frequency of the vehicle traveling along the given c to locations l_i and l_j, respectively.

Then, the optimal ranking $l_i >_{v,t} l_j$ can be addressed by maximizing the following posterior probability:

$$\begin{aligned} &\arg\max_{\Theta} p\left(\Theta \mid \widehat{l}_{v,c,l_i} > \widehat{s}_{v,c,l_j}\right) \\ &\propto \arg\max_{\Theta} \left(\widehat{l}_{v,c,l_i}\Theta > \widehat{l}_{v,c,l_j}\right) p(\Theta). \end{aligned} \tag{9}$$

By assuming all the vehicles are independent of each other, $p(\Theta)$ is a normal distribution with zero mean and a variance-covariance matrix $\lambda_\Theta I$, that is, $p(\Theta) \sim N(0, \lambda_\Theta I)$. Meanwhile, we adopt the logistic sigmoid $\sigma(x) \coloneqq 1/(1 + e^{-x})$ to approximate the likelihood of vehicle's preference over l_i and l_j and utilize $C_{\langle v,i,j\rangle}$ to measure the confidence

$$\begin{aligned} &p\left(\widehat{l}_{v,c,l_i}\Theta > \widehat{l}_{v,c,l_j}\right) \\ &= \sigma\left(\widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}\right) = \frac{1}{1 + e^{-C_{\langle v,i,j\rangle}\left(\widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}\right)}}. \end{aligned} \tag{10}$$

Furthermore, the alternating maximum a posteriori estimation in logarithmic scale is calculated as follows:

$$\begin{aligned} &\arg\max_{\Theta} \prod_{(v,c,l)\in V\times C\times L} \sigma\left(\widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}\right) p(\Theta) \\ &= \arg\max_{\Theta} \ln\left(\prod_{(v,c,l)\in V\times C\times L} \sigma\left(\widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}\right) p(\Theta)\right) \\ &= \arg\max_{\Theta} \sum_{(v,c,l)\in V\times C\times L} \ln \sigma\left(\widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}\right) \\ &\quad - \lambda_\Theta \|\Theta\|_F^2. \end{aligned} \tag{11}$$

Then, the stochastic gradient descent algorithm is used to optimize the above objective function. Once the parameter set $\Theta = \{V, C, L^V, L^C\}$ is acquired, the third-rank tensor of transition preference over locations can be recovered. The probability of different next location for each vehicle can then be obtained. The complete process of our proposed prediction model is detailed in Algorithm 1.

5. Security Analysis

In this section, we present the prediction accuracy, the security analysis, and their relation.

5.1. A Discussion of Predication Accuracy. Here, we discuss the way of lowering the prediction accuracy and make it more difficult for the adversary to grasp vehicle's movement and launch an attack.

Given the obtained next location prediction list N for a target taxi v^*, we first analyze the factors of affecting predict accuracy. As shown in Figure 4, there is a vehicle at location A. If we only know its current location, there will be four choices to predict. The next location is to be predicted based on the vehicle's current location, which can be formalized

(1) **Input:** Data D
(2) **for** $v \subset V$, $l \subset L$ **do**
(3) $C \leftarrow k$-order Markov chain (L)
(4) **end for**
(5) draw $\widehat{\vec{V}}$, $\widehat{\vec{C}}$, $\widehat{\vec{L}}^{V}$, $\widehat{\vec{L}}^{C}$ from $\mathcal{N}(0, \lambda_{\Theta} I)$
(6) **repeat**
(7) draw (v, c, i, j) form D
(8) $T_i \leftarrow$ frequency from c to location i
(9) $T_j \leftarrow$ frequency from c to location j
(10) **if** $T_i < T_j$ **then**
(11) swap l_i with l_j
(12) **end if**
(13) $C_{\langle v,i,j \rangle} = 1 - (T_i - T_j)/(T_i + T_j)$
(14) $\widehat{l}_{v,c,l_i,l_j} \leftarrow \widehat{l}_{v,c,l_i} - \widehat{l}_{v,c,l_j}$
(15) $\delta \leftarrow (1 - \sigma(C_{\langle v,i,j \rangle} \widehat{l}_{v,c,l_i,l_j}))$
(16) **for** $f = 1$ to k **do**
(17) $\widehat{v}^{L}_{v,f} \leftarrow \widehat{v}^{L}_{v,f} + \alpha(C_{\langle v,i,j \rangle} \delta(\widehat{l}^{V}_{l_i,f} - \widehat{l}^{V}_{l_j,f}) - \lambda_{\theta} \widehat{v}^{L}_{v,f})$
(18) $\widehat{c}^{L}_{c,f} \leftarrow \widehat{c}^{L}_{c,f} + \alpha(C_{\langle v,i,j \rangle} \delta(\widehat{l}^{C}_{l_i,f} - \widehat{l}^{C}_{l_j,f}) - \lambda_{\theta} \widehat{c}^{L}_{c,f})$
(19) $\widehat{l}^{V}_{l_i,f} \leftarrow \widehat{l}^{V}_{l_i,f} + \alpha(C_{\langle v,i,j \rangle} \delta \widehat{v}^{L}_{v,f} - \lambda_{\theta} \widehat{l}^{V}_{l_i,f})$
(20) $\widehat{l}^{V}_{l_j,f} \leftarrow \widehat{l}^{V}_{l_j,f} + \alpha(-C_{\langle v,i,j \rangle} \delta \widehat{v}^{L}_{v,f} - \lambda_{\theta} \widehat{l}^{V}_{l_j,f})$
(21) $\widehat{l}^{C}_{l_i,f} \leftarrow \widehat{l}^{C}_{l_i,f} + \alpha(C_{\langle v,i,j \rangle} \delta \widehat{c}^{L}_{c,f} - \lambda_{\theta} \widehat{l}^{C}_{l_i,f})$
(22) $\widehat{l}^{C}_{l_j,f} \leftarrow \widehat{l}^{C}_{l_j,f} + \alpha(-C_{\langle v,i,j \rangle} \delta \widehat{c}^{L}_{c,f} - \lambda_{\theta} \widehat{l}^{C}_{l_j,f})$
(23) **end for**
(24) **until** convergence **Output**: $\widehat{\vec{V}}$, $\widehat{\vec{C}}$, $\widehat{\vec{L}}^{V}$, $\widehat{\vec{L}}^{C}$

Algorithm 1: Our proposed tensor-based prediction model.

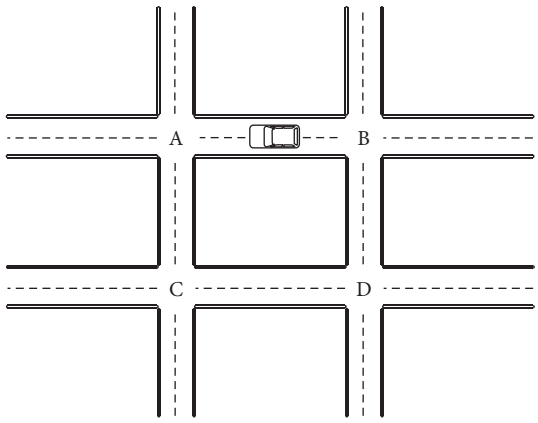


Figure 4: A snapshot to illustrate how road structure affects the prediction accuracy.

using the 1-order Markov chain. However, if we know that this vehicle came from location B, then it infers that we are less likely to come back to B. With the sequence (B, A), we can find out the next location is most likely to appear in the next three locations, which can be formalized using the 2-order Markov chain. Furthermore, if we extend the case to the 3-order chain, knowing this vehicle previous driving sequence is (D, B, A), then the probability that it comes back to location B based on the 2-order Markov chain will be further lowered. And the vehicle has less chance to go to location C as going to C will make a detour. In general, we argue that the prediction accuracy is affected by the order of Markov chain. Protecting location privacy means to lower the prediction accuracy for the adversary. So we propose a strategy to prevent the adversary from acquiring the consecutive location information to form high order Markov chain. As the vehicle's average speed and average length between two consecutive intersections can be speculated in the city, thus a time slot in which one cannot query the vehicle about the location information can be deducted.

5.2. Security Analysis for Goals. The security of k-order FPMC strategy is analyzed in Theorem 2.

Theorem 2. *The (p, t) customers' location privacy is preserved in the k-order FPMC strategy, when the successful prediction rate of the k-order FPMC strategy in time $f(1)$ is less than p, where $f(k)$ is a function which indicates the average time for vehicles passing k consecutive location points.*

Proof. Intuitively, in the k-order FPMC strategy, since the vehicular location is reported to the center in time $f(k)$, the trajectory during the time $f(k-1)$ to $f(k)$ cannot be predicted to the inside attacker, according to the ability of inside attacker defined in the adversarial model. Therefore, the customers' location privacy is preserved. Furthermore, if

the inside attacker uses the discontinuous customer's location information, the successful prediction rate is p, due to the assumption in the theorem.

We further use the mathematical induction to formally demonstrate the concrete probability that the adversary $\mathscr{A}$ obtains the user's trajectory in time $f(k)$.

(1) $\Pr[\mathscr{A}_1]$ is less than p, due to the assumption.

(2) Assume that $\Pr[\mathscr{A}_t]$ represents the successful prediction rate of $\mathscr{A}$ in time $f(t)$. Therefore, $\Pr[\mathscr{A}_{k-1}] = p$ means that the adversary computes the user's route in time $f(k-1)$ as p due to $f(k-2)$. Since all the successful prediction rate is the same from the viewpoint of the adversary, the successful prediction rate $\Pr[\mathscr{A}_k]$ of $\mathscr{A}$ in time $f(k)$ is

$$\begin{aligned}\Pr\left[\mathscr{A}_k\right] &= \Pr\left[\mathscr{A}_{k-1}\right] = \Pr\left[\mathscr{A}_{k-2}\right] = \cdots = \Pr\left[\mathscr{A}_1\right] \\ &\leq p.\end{aligned} \tag{12}$$

Compared with Definition 1, the (p,t) customers' location privacy is preserved in the k-order FPMC strategy. In summary, the k-order FPMC strategy achieves the security goal, as long as $\Pr[\mathscr{A}_1]$ is less than p. □

Note that $\Pr[\mathscr{A}_1]$ is negligible, because the user's historical data cannot be obtained by the inside attack from the very beginning. Thus, the user's route is privacy-preserving for the inside attack in the k-order FPMC strategy.

The rest is to quantitatively measure the successful prediction rate p and the time interval t of the k-order FPMC strategy. More technical details for the parameters p and t are demonstrated in Section 6.

6. Experimental Analysis

In this section, we first present the details of the dataset used in the experiments. Then, we evaluate our proposed method and compare the model performance with several competitive prediction algorithms. The experimental results show that our approach significantly outperforms the-state-of-art approaches. It is possible for the adversary to utilize such sophisticated approach to launch the location attack successfully. We also provide the strategy and analyze the possibilities of reducing such risks.

6.1. Dataset Description. In this paper, we adopt the real traces and traffic statistics in Shanghai SUVnet [36]. We study the V2V network formed by more than 4,000 taxies in Shanghai, which are monitored by SUVnet. The GPS locations of all taxies are periodically collected. The updating time interval is around 30 seconds.

A 12.6 km × 12.9 km region in Shanghai is used as our test-bed. As shown in Figure 5, the highlighted streets on the MapInfo-empowered map form a road network graph. The statistics of the datasets being used in our experiments are tabulated in Table 1. In all our experiments, we use four-day data (S_{train}) for training and the next one-day data (S_{test}) for testing. The algorithm was trained with S_{train} and then the performance is measured on S_{test}.

Figure 5: A snapshot of selected areas road network in Shanghai.

Table 1: The statistics of the Shanghai SUVnet datasets.

Taxi number	Node	Road	Days
645	124	211	5

6.2. Evaluation Metric. Recall that the prediction task is to provide a list of predicted locations among which only at most one will be picked by the researcher or the adversary. This will make the prediction precision evaluated by the conventional precision metric to be lower than 1/|list|. Instead, we adopt a precision metric for our particular prediction task, and we divide the test data into 24 parts according to the hours of the day and perform the proposed method for each hour and then take the average as the final results; the proposed metric is given as

$$P@N = \frac{1}{|V|}\frac{1}{|T|}\sum P_u@N = \sum \frac{1}{|V * T|}\sum \frac{S_v}{N_v}, \tag{13}$$

where $|V|$ denotes the number of the vehicles and S_v and N_v represent the counts of correct predictions and the total number of recommendation rounds for each vehicle, respectively.

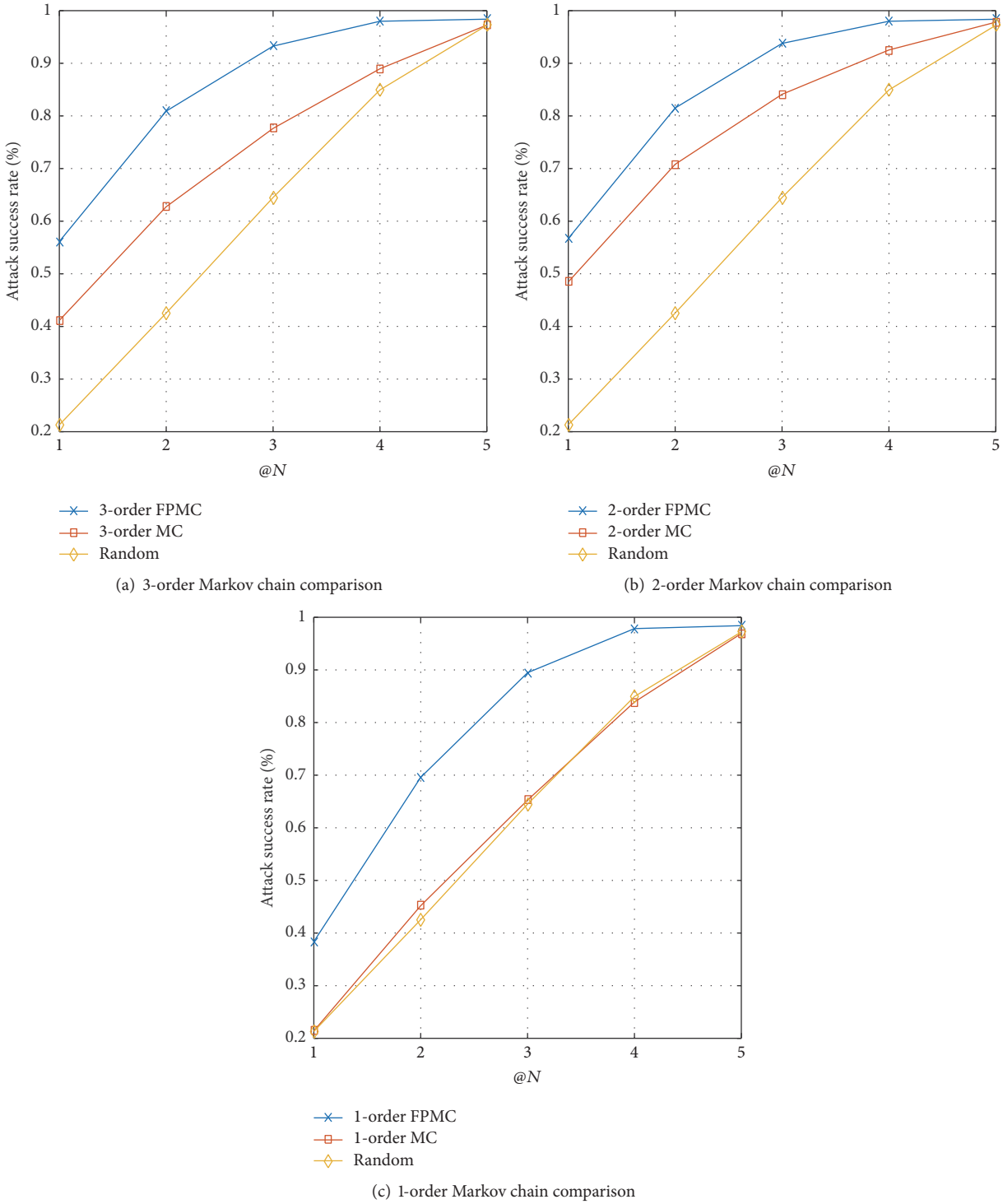
6.3. Performance Comparison. In this section, we compare our method with the following state-of-the-art alternatives:

(1) *MC*. Conventional Markov chain is widely adopted in many prediction tasks including location prediction [12, 19, 23–26]. In this paper, we apply the conventional Markov chain to predict the next location for comparison.

(2) *ANN*. Artificial Neural Networks (ANN) consider the next place prediction as a classification task, given the current place, and the other features, that is, time, to output the predicted next location [37, 38]. We apply ANN in this paper as a none Markov model baseline.

(3) *Random*. This method is a random guess on the set of adjacent intersections of vehicle's current location.

6.4. Experimental Results. To evaluate the prediction model, we first compare our proposed method with the conventional



(a) 3-order Markov chain comparison

(b) 2-order Markov chain comparison

(c) 1-order Markov chain comparison

Figure 6: Prediction comparison under different orders of Markov chain.

Markov chain approach and random guess. The experimental results are shown in Figure 6. Each subfigure of Figure 6 represents the obtained results set with different settings of the order of Markov chain. It is obvious that our proposed k-order FPMC approach outperforms the conventional Markov chain-based model and the random approach. The main reason behind is that the tensor-based approach is capable of capturing the spatiotemporal regularities of the vehicles' traces better and capturing their resemblance simultaneously.

From Figure 6, we can also observe that when N (the size of the candidate location set) goes to 5, the precision

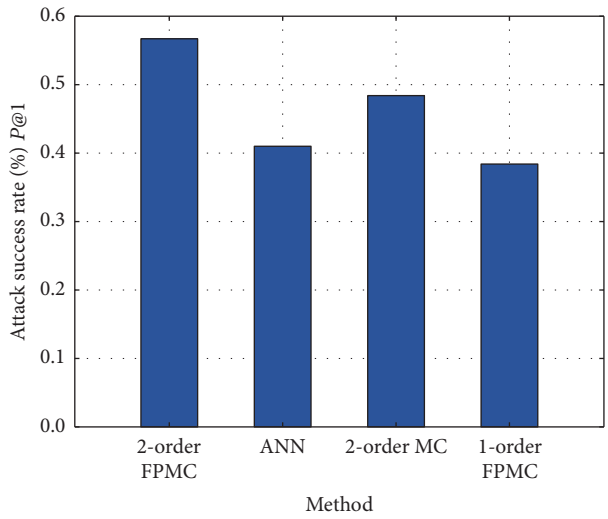


Figure 7: Performance comparison under different prediction models.

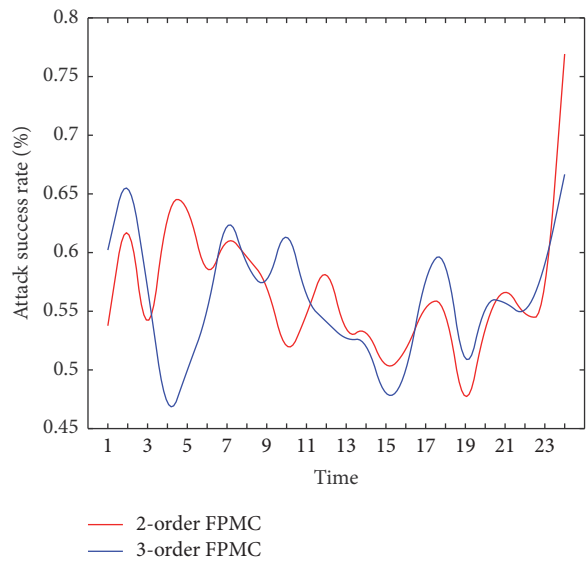


Figure 8: Attack success rate versus time.

almost reached 100%. Actually, one intersection usually has 4 adjacent intersections on average, and the result is in line with the actual situation. However, when we talk about location privacy problem, what we are concerned about is mainly the result of $P@1$. The adversary takes the most likely location from the prediction list as the target location. From the results of $P@1$, we can conclude that the localization attack using our proposed sophisticated method is a serious threat.

In order to reflect how the order of Markov chain k influences the prediction accuracy in our framework, we conduct a comparison for different setting k (varying from 1 to 3). As shown in Figure 9(a), we can see that the 2-order and 3-order one almost have the same prediction accuracy, but when the order deceases to 1, the location prediction accuracy drops sharply. There is nearly 20% disparity between the 2-order and 1-order chains, which demonstrates our conjecture as illustrated in Section 5.1. And we could decrease the risk of the location attacks by carefully concealing the consecutive location information and making the information unable to form a high order Markov chain.

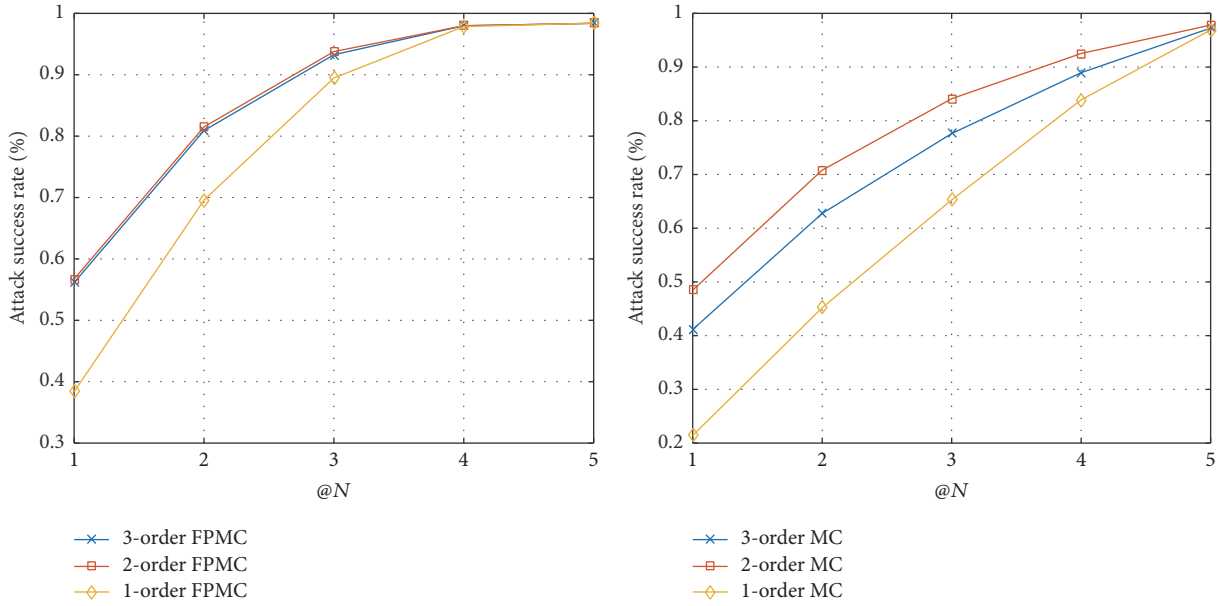
We also conduct the experiments by using ANN a chic prediction model for prediction comparison. From Figure 7, we can observe that the 2-order Markov chain FPMC have the best performance in terms of the prediction accuracy $P@1$, which further manifest the advantage of our proposed prediction method. The 1-order Markov chain FPMC have the worst performance which demonstrates that our t-limit model can achieve the security goal.

In order to distinguish the experimental results obtained by FPMC and MC with different order settings, we represent the results in Figures 9(a) and 9(b), respectively. We can observe that 2-order MC performs better than 3-order MC, while 2-order FPMC performs nearly the same as 3-order FPMC. Figure 8 illustrates a fine-grained comparison along with the time. We argue that 3-order FPMC and 3-order MC can catch the road structure constraints better than 2-order approaches; however the historical data of 3-order FPMC is sparse which leads to the less pleasant results for MC. Nevertheless, for tensor-based approach FPMC, we can infer a good performance while the historical data is sparse.

To suggest a proper setting of the protection time interval t, we investigate the average passing time on the road sections. We calculate the passing time between every road segment at different time period on each day and calculate the variance during each day. Table 2 tabulates the detailed time variance for 24 time periods. From the table we can see that the average passing time tends to be stable, and the difference is no more than 90 s. Therefore, it is reasonable for us to say that the average road passing time indicates a good time interval setting to protect location privacy. During the time interval, we suggest vehicle not to release its location; thus the 2-order Markov chain can not be formed easily which will restrict the performance of most of existing location prediction models and thus reduce the risk of the localization attacks.

7. Conclusion

In this paper, we proposed a novel approach to predict vehicle's next location and a strategy to protect vehicle location privacy from such model. First, BPR-based k-order Markov chain tensor model is proposed to predict the location preference by exploring the Markov property of the taxis' travel trajectory found in datasets. Then, based on the road structure we argue that the order of Markov chain k plays an important role to affect the prediction accuracy. Therefore, we analyzed the average road segment passing time and suggest breaking the continuity of the passing locations during the time period to prevent the adversary from forming a high order Markov chain. The proposed approach carefully combines the factors of spatial and temporal information. Performance evaluations are conducted based on the Shanghai SUVnet dataset which shows that the proposed approach can improve the prediction accuracy significantly compared to several existing state-of-the-art methods, and



(a) FPMC comparison

(b) The conventional MC comparison

Figure 9: Prediction comparison.

Table 2: The statistics of the road segment passing time.

Time	Time variance (s^2)
0	4440.0
1	5692.0
2	5901.0
3	7964.0
4	9288.0
5	7926.0
6	5950.0
7	4208.0
8	3003.0
9	3928.0
10	4330.0
11	4051.0
12	4146.0
13	5421.0
14	3784.0
15	3594.0
16	3629.0
17	2826.0
18	2867.0
19	3018.0
20	3742.0
21	3722.0
22	3085.0
23	2489.0

our proposed privacy protection schema is feasible to prevent leaking location privacy from such sophisticated predication model.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would also like to thank Dr. Xiaofeng Zhang affiliated with Department of Computer Science, Graduate School of Harbin Institute of Technology, China, for his great effort of helping us further improve the quality of the paper. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant nos. 61300178 and 61300177 and the National Program on Key Basic Research Project under Grant no. 2013CB329605.

References

[1] D. Anthony, T. Henderson, and D. Kotz, "Privacy in location-aware computing environments," *IEEE Pervasive Computing*, vol. 6, no. 4, pp. 64–72, 2007.

[2] G. Hosein, "They know where you are," *Index on Censorship*, vol. 36, no. 4, pp. 132–136, 2007.

[3] L. Barkhuus and A. K. Dey, "Location-based services for mobile telephony: a study of users' privacy concerns," in *Proceedings of the International Conference on Human-Computer Interaction (INTERACT '03)*, pp. 709–712, Zurich, Switzerland, September 2003.

[4] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: a survey," *IEEE Vehicular Technology Magazine*, vol. 2, no. 2, pp. 12–22, 2007.

[5] E. Koukoumidis, L.-S. Peh, and M. R. Martonosi, "SignalGuru: leveraging mobile phones for collaborative traffic signal schedule advisory," in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys '11)*, pp. 127–140, Washington, DC, USA, June 2011.

[6] I. Bilogrevic, K. Huguenin, M. Jadliwala et al., "Inferring social ties in academic networks using short-range wireless communications," in *Proceedings of the ACM Workshop on Workshop on Privacy in the Electronic Society*, pp. 179–188, 2013.

[7] L. Liao, D. Fox, and H. Kautz, "Extracting places and activities from GPS traces using hierarchical conditional random fields," *International Journal of Robotics Research*, vol. 26, no. 1, pp. 119–134, 2007.

[8] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[9] M. Gruteser and B. Hoh, "On the anonymity of periodic location samples," in *Proceedings of the 2nd International Conference on Security in Pervasive Computing (SPC '05)*, pp. 179–192, Boppard, Germany, April 2005.

[10] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS '07)*, pp. 161–171, Alexandria, Va, USA, October 2007.

[11] J. T. Meyerowitz and R. R. Choudhury, "Realtime location privacy via mobility prediction: creating confusion at crossroads," in *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications (HotMobile '09)*, pp. 1–6, Santa Cruz, Calif, USA, February 2009.

[12] Y. Wu, Y. Zhu, and B. Li, "Trajectory improves data delivery in vehicular networks," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM '11)*, pp. 2183–2191, Shanghai, China, April 2011.

[13] D. J. Patterson, L. Liao, D. Fox, and H. Kautz, *Inferring High-Level Behavior from Low-Level Sensors*, Springer, Berlin, Germany, 2003.

[14] J. Krumm, J. Letchner, and E. Horvitz, "Map matching with travel time constraints," in *Intelligent Transportation Systems*, pp. 1–1102, 2007.

[15] J. Krumm and E. Horvitz, "Predestination: inferring destinations from partial trajectories," in *Proceedings of the Ubiquitous Computing, International Conference (UBICOMP '06)*, pp. 243–260, Orange County, Calif, USA, September 2006.

[16] Y. Matsuo, N. Okazaki, K. Izumi et al., "Inferring long-term user properties based on users' location history," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 2159–2165, Hyderabad, India, January 2007.

[17] S. A. Franken, "The location privacy protection act of 2011(s. 1223)," *Bill Summary*, 2013.

[18] K. Minami and N. Borisov, "Protecting location privacy against inference attacks," in *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS '10)*, pp. 711–713, Chicago, Ill, USA, October 2010.

[19] R. Shokri, G. Theodorakopoulos, G. Danezis, J. P. Hubaux, and J. Y. L. Boudec, *Quantifying Location Privacy: The Case of Sporadic Location Exposure*, Springer, Berlin, Germany, 2011.

[20] J. Krumm, "Where will they turn: predicting turn proportions at intersections," *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 591–599, 2010.

[21] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the International Workshop on Trajectory Data Mining and Analysis (TDM) in Conjunction with the 13th International Conference on Ubiquitous Computing (TDMA '11)*, pp. 23–30, Beijing, China, September 2011.

[22] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior," in *Proceedings of the 10th ACM International Conference on Ubiquitous Computing (UbiComp '08)*, pp. 322–331, Seoul, South Korea, September 2008.

[23] J. Qin, H. Zhu, Y. Zhu, L. Lu, G. Xue, and M. Li, "Post: exploiting dynamic sociality for mobile advertising in vehicular networks," *IEEE Transactions on Parallel Distributed Systems*, vol. 27, no. 6, pp. 1770–1782, 2016.

[24] S. Chen, Y. Li, W. Ren, D. Jin, and P. Hui, "Location prediction for large scale urban vehicular mobility," in *Proceedings of the 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC '13)*, pp. 1733–1737, Sardinia, Italy, July 2013.

[25] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (WPES '08)*, pp. 23–32, ACM, 2008.

[26] S. Gambs, M.-O. Killijian, and M. N. D. P. Cortez, "De-anonymization attack on geolocated data," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1597–1614, 2014.

[27] Y. Zheng, X. Xie, and W. Y. Ma, "Geolife: a collaborative social networking service among user, location and trajectory," *Bulletin of the Technical Committee on Data Engineering*, vol. 33, no. 2, pp. 32–39, 2010.

[28] Z. Zhang, L. Zhu, L. Liao, and M. Wang, "Computationally sound symbolic security reduction analysis of the group key exchange protocols using bilinear pairings," *Information Sciences*, vol. 209, no. 22, pp. 93–112, 2012.

[29] Z. Zhang, C. Jin, M. Li, and L. Zhu, "A perturbed compressed sensing protocol for crowd sensing," *Mobile Information Systems*, vol. 2016, Article ID 1763416, 9 pages, 2016.

[30] Z. Zhang, Z. Qin, L. Zhu, J. Weng, and K. Ren, "Cost-friendly differential privacy for smart meters: exploiting the dual roles of the noise," *IEEE Transactions on Smart Grid*, 2016.

[31] S. Rendle and L. Schmidt-Thieme, "Pairwise interaction tensor factorization for personalized tag recommendation," in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pp. 81–90, ACM, New York, NY, USA, February 2010.

[32] C. Cheng, H. Yang, M. R. Lyu, and I. King, "Where you like to go next: successive point-of-interest recommendation," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 2605–2611, Beijing, China, August 2013.

[33] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI '09)*, UAI, pp. 452–461, AUAI Press, Montreal, Canada, June 2009.

[34] W. Pan, H. Zhong, C. Xu, and Z. Ming, "Adaptive Bayesian personalized ranking for heterogeneous implicit feedbacks," *Knowledge-Based Systems*, vol. 73, pp. 173–180, 2015.

[35] S. Wang, X. Zhou, Z. Wang, and M. Zhang, "Please spread: recommending tweets for retweeting with implicit feedback," in *Proceedings of the ACM Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media (DUBMMSM '12)*, pp. 19–22, Maui, Hawaii, USA, October 2012.

[36] S. J. University, Traffic information grid term, grid computing cernter. Shanghai taxi trace data, http://wirelesslab.sjtu.edu.cn/.

[37] V. Etter, M. Kafsi, and E. Kazemi, "Been there, done that: what your mobility traces reveal about your behavior," in *Proceedings*

of the Mobile Data Challenge by Nokia Workshop, in Conjunction with International Conference on Pervasive Computing, EPFL-CONF-178426, pp. 1–6, Newcastle, UK, June 2012.

[38] J. V. Subramanian and M. A. K. Sadiq, "Implementation of artificial neural network for mobile movement prediction," *Indian Journal of Science and Technology*, vol. 7, no. 6, pp. 858–863, 2014.