Study on Personalized Location Privacy Preservation Algorithms Based on Road Networks

Hongyun Xu, Jun Yang, Yong Zhang(✉), Mengzhen Xu, and Jiayi Gan

School of Computer Science and Engineering,
South China University of Technology, Guangzhou 510006, China
z.yoo@qq.com

Abstract. It is very important for LBS popularization and application to study personalized location privacy preservation algorithms based on road networks for the mobile users. This paper proposes the Prediction Group by L algorithm (i.e., PL) in which the road networks is represented as a weighted graph, and the value of the weight of each edge is equal to its selection rate; the edges of the graph are sorted by the depth-first search algorithm, and grouped by the privacy degree of user where the group is used as the anonymous edge set (i.e., AES) to realize the location privacy preservation. The experimental results show that PL has high success rate of privacy. Additionally, it is able to provide higher quality personalized location privacy preservation because the AES generated by this algorithm is more approached to the privacy requirements of users than some other typical algorithms.

Keywords: Road networks · Location privacy · K-anonymity · L-diversity · Location-based service

1 Introduction

With the popularization of smart phones and portable devices, location-based service (i.e., LBS) applications attract more and more users. Typical LBS applications include the nearest point of interest (POI) query, location-aware advertisement, and road navigation, etc.

However, if users want to obtain LBS, they have to send their private information (such as location, identity) to location server (i.e., LS), which is often untrusted and may reveal users' information, thus threatening user's privacy [1, 2]. Therefore, LBS applications should consider privacy preservation of user's information. In recent years, there have been lots of researches on how to preserve user's privacy information. So far, the mainstream methods on privacy preservation in LBS can be classified into two kinds.

The first one is based on Euclidean Space, which supposes that the user can move in a space without limit. This kind mainly uses K-anonymity [3, 4] and obfuscation region [5]. K-anonymity guarantees there are at least other K-1 users in the obfuscation region, and user's practical location will be replaced with the obfuscation region and

G. Wang et al. (Eds.): ICA3PP 2015 Workshops, LNCS 9532, pp. 35–45, 2015.
DOI: 10.1007/978-3-319-27161-3_4

sent to server. For personalized privacy preservation, this kind provides different degrees of privacy preservation [9, 10] by means of adjusting parameter K.

The second one is based on road networks, and it assumes that mobile user and query object are both located on practical roads. This kind [6–8] transforms the road network to graph, and uses K-anonymity and L-diversity [11] to implement privacy preservation, where L-diversity guarantees that there are at least L different physical locations in the obfuscation region. For personalized preservation, this kind completes the preservation by adjusting both parameters K and L of the obfuscation region.

The second one above can provide a higher practical value than the first one, since it considers the practical situations. Therefore we do some researches on the second method and propose a new personalized location privacy preservation algorithm. The main contributions of our paper are listed below:

- Based on the amount of mobile users on the road networks and the length of road sections, we define the road prosperous degree and selection rate, which are used for measuring probability of road section that users locates at.
- By using the analysis of Markov predicting method, we draw the conclusion that when mobile user is at the cross road, the larger selection rate of the road section is, the higher probability the user drives in.
- According to road selection rate and road prosperous degree, we propose a personalized location privacy preservation algorithm which satisfies both K-anonymity and L-diversity of user requirements.
- After lots of experiments and comparisons between our algorithm and some other typical algorithms, we make some analysis and draw a conclusion that our algorithm provides higher quality for personalized privacy preservation than other algorithms.

In Sect. 2 we summarize the related works; Sect. 3 introduces the road networks model, query model, system structure and attack model; Sect. 4 describes the algorithm in detail; Sect. 5 contains theory analysis and experiment analysis; and Sect. 6 is the conclusion of this paper.

2 Related Works

There have been many researches on the location privacy preservation problems based on road networks. Mouratidis et al. [7] propose using depth-first search algorithm (i.e., DFS algorithm) and breadth-first search algorithm (i.e., BFS algorithm). They use DFS or BFS to implement the process of linearization of road networks, and then generate the obfuscation region to implement the privacy preservation. However, this method doesn't consider L-diversity, which may have a hidden threat.

Wang et al. [6] propose a location privacy preservation algorithm called XStar. It represents the road section as a star in the intersection and it uses star and its neighbor stars to implement the anonymity. This algorithm considers both K-anonymity and L-diversity, however, its success rate is low.

Xue et al. [8] propose location privacy preservation method called anonymous Cycle and Forest (i.e., CCF). It transforms road networks to graph and uses BFS to find some satisfactory cycles or forests as an AES. However, if privacy degree of user is relatively low, the AES generated by this algorithm may contain more users and road sections than user's expectation, therefore it may result in low query quality.

To solve the above problems, we propose the concepts of road prosperous degree and selection rate in road networks, and design a personalized location privacy preservation algorithm which is based on road selection rate. Our algorithm, called Prediction Group by L algorithm, satisfies both user's K-anonymity and L-diversity.

3 System Models

This section introduces some system models, including the road networks model, the query model, system structure and attack model.

3.1 Road Networks Model

The road networks can be modeled as an undirected graph G = (V, E), where V represents nodes set, and E represents road sections set. The degree of nodes represents the amount of road sections which are connected with the intersections. As shown in Fig. 1, we can see that it can be represented as an undirected graph G = (V, E), and the nodes set V = {n1, n2, n3…, n13}, road sections set E = {n1n2, n2n3, n3n4, n3n12…, n12n13, n13n4}, the degree of node n3 is 4.

3.2 Query Model

We formulate the model Qu = {id, loc, req, K, L} as user's query model, where Qu represents a query from user u, and it is composed of user's identification information id, location information loc, query requirement req, location privacy requirements K and L. The parameters K and L represent K-anonymity and L-diversity respectively.

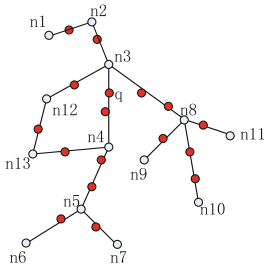


Fig. 1. Road networks

Fig. 2. Central server structure

3.3 Location Privacy System Structure

We use Central Server Structure to preserve location privacy, as shown in Fig. 2, and we add a Location Privacy Server (i.e., LPS) in the middle. The main function of LPS is to receive query and generate AES according to parameters K, L and loc from Qu.

Suppose the AES is loc', the identification information of users in AES is id', combine loc', id' and req, LPS can get the new query Qu' = {id', loc', req}, then it sends Qu' to LBS server. Next LBS server returns the query results (called candidate set) to LPS, and then LPS filters the candidate sets and send the appropriate results to users.

3.4 Attack Model

The goal of attackers is to acquire users' privacy information from LPS. However, the ability of an attacker is related with his background and inference algorithms.

We suppose that the attacker knows the structure of road networks and the distribution of users in the networks. The attacker can take use of distribution to calculate the probability of each AES that the users locate at, and attack the users in the road section with higher probability since it has higher success rate.

Additionally, suppose attacker knows the AES of users, the topological graph of road networks, and the personalized privacy algorithm which is used to generate the AES. The attacker can implement location privacy preservation algorithm for each edge of AES and attack the users in the edge with higher probability.

4 Algorithm Design

This section introduces some related definitions and designs for the personalized location privacy preservation algorithm, Prediction Group by L.

4.1 Related Definitions

Road Prosperous Degree. It represents the amount of mobile users in a unit length of a road section. We use W to represent it. From the definition, we know if the road section has higher prosperous degree, then it has more mobile users in unit length of this road section.

As shown in Fig. 1, the red nodes represent mobile users, and we suppose the length of n3n8, n8n11, n8n10, n8n9 are all equal to 1. Then the road prosperous degrees of them are W(n3, n8) = 2, W(n8, n11) = 1, W(n8, n10) = 2, W(n8, n9) = 1 respectively.

Selection Rate. It defines the probability which road section the user will drive in and it is calculated by the ratio of the prosperous degree of a road section and the sum of all prosperous degrees of connected road sections. We use P to represent it. Usually it is used for the case when mobile user is in the cross road (like node n8 in Fig. 1), and it can infer most possible road section that user locates at.

As shown in Fig. 1, suppose the user locates at n8, then P(n3, n8) = W(n3, n8) /(W (n3, n8) + W(n8, n11) + W(n8, n10) + W(n8, n9)) = 1/3, in the same way, we can get that P(n8, n11) = 1/6, P(n8, n10) = 1/3, P(n8, n9) = 1/6, etc.

Open Node [7]. It represents the node in the AES which satisfies that at least one road section connected with this node is not in the AES.

KNN Query of Anonymous Edge Set [7]. It represents doing K-Nearest Neighbor query for the open node in the AES, namely, retrieves K objects that are nearest to the node, and query all the objects in the AES.

4.2 Prediction-Based Grouping Algorithm

4.2.1 Markov Prediction Analysis [12]

When the user drives in the intersection of the road networks, what is the characteristic of the next road section driving in? We use Markov Prediction method to make some analysis.

Markov Prediction method is to determine the future status according to the original probability and status transition probability of different status of an event. The probability of event transition is defined by Eq. (1), and we choose the biggest probability to determine the future status, namely choose the biggest one in P.

$$P = \begin{bmatrix} P11 & P12 & \dots & P1n \\ P21 & P22 & \dots & P2n \\ \dots & \dots & \dots & \dots \\ Pn1 & Pn2 & \dots & Pnn \end{bmatrix} \tag{1}$$

Furthermore, we use selection rate as the transition probability of Markov Prediction method, which is used to predict the next road section that the mobile user will choose to drive in. And this analysis is based on the road networks of Oldenburg in Germany [13].

We set the number of mobile users N1 in the road networks to 6329 at time t1, and then use the biggest selection rate and the smallest selection rate as the prediction results of status transition probability; next we can get the result as Table 1 shown. In Table 1, the first column means in the next time t2, the number of users that appear in the predicted road section is N2; the second column means percentage of the users who are driving to the predicted road section. We use PS to represent it. The first row means the predicted result with the biggest selection rate; and the second row means that with the smallest selection rate. From Table 1, we can see that use bigger selection rate can have higher success rate than use smaller selection rate in predicting.

Table 1. Results of Markov prediction

	N2	PS
P set to largest	2178	34.41
P set to smallest	696	11.00

From above all, we can see that the bigger the road selection rate is, the higher probability that the mobile user chooses this road section to drive in; otherwise, the lower probability that user choose this section to drive.

4.2.2 PL Algorithm

Prediction Group by L algorithm (i.e., PL), uses two privacy parameters, K-anonymity and L-diversity, to realize the personalized location privacy preservation of users. First it uses selection rate as the weight of road section, and depth-first search algorithm to traverse the road networks according to the value of weight, to get the linearized sequence. Next the algorithm groups the sequence according to user's personalized privacy parameters K and L, which makes the distribution of mobile users in different roads uniform. In this way, it can decrease the success rate of inference attack to weight of edges. The main steps of PL algorithm are shown in Algorithm 1.

Algorithm 1. Prediction Group by L Algorithm
Input: Road network G, the distribution of mobile users in G, the location information of user u, the requirement of privacy parameters K and L
Output: Anonymous edge set S
Step 1, calculate the prosperous degree and selection rate of each road section in G;
Step 2, use DFS algorithm to sort each road section in G, linearize G and get the linearized Array according to selection rate P. Obtain the index number IdL of Array where the user u locates in;
Step 3, group Array by user's requirement for L-diversity, and set the starting index number StartId = (IdL / L) * L and the ending index number EndId = (IdL / L) * L + L - 1 to S;
Step 4, calculate the number of mobile users in S, mark it as k';
Step 5, if k' < K, L adds 1, and jump to step 3;
Step 6, output S, end.

Here we give an example to show how to get the satisfied S. Suppose IdL = 27, L = 5, then StartId = (27 /5) * 5 = 25, EndId = (27 /5) * 5 + 5 - 1 = 29; calculate the users in S, if k' < K, L adds 1, and we get new StartId = (27 /6) * 6 = 24 and EndId = (27 /6) * 6 + 6 - 1 = 29. Repeat these steps, we can gradually increase the size of S and finally get the satisfied S.

5 Simulation Experiments and Analysis

This section compares and analyzes our algorithm and typical algorithms by experiments. The experiments are realized by Java, and the coding environment is Eclipse. The hardware environment is Intel(R) Pentium(R) 4 CPU 2.66 GHz, 1.49 GB internal memory. The operating system is Microsoft Windows XP SP3. The experiment platform uses the famous road networks generating platform [13], the Network-based Generator of Moving Objects (i.e., NGMO). It uses the map of Oldenburg in Germany to simulate experiments.

5.1 Parameters Setting

The parameters setting of experiments are shown in Table 2. The amount of nodes, edges and average length of edges represent the corresponding parts in the map, and the query amount represents the amount of queries that users send. User amount represents the amount of mobile users in the road network, and POI amount represents the total amount of POIs in the road network. Additionally, the value of K, L, the maximal L and *k*NN satisfy normal distribution.

Table 2. Simulation environment parameter table

Item	Values	Item	Values
Nodes	6105	Minimal K	3
Total edges	7035	Average L	5
Average edges length	184	Minimal L	3
Query amount	10000	Maximal L	20
User amount	17845	*k*NN	5
Average K	5	POI	4949

5.2 Algorithm Evaluation Target

In our paper we evaluate the algorithm in four aspects, namely the average information entropy, average anonymous edge set, average query cost and privacy success rate.

$$\text{Entropy} = -\sum_{l \in S} P(l)*\log(P(l)) \tag{2}$$

Average Information Entropy. It defined by Eq. (2) [6]. In Eq. (2) $P(l)$ represents the probability that user is in the road section l of AES. When the amount of edges is steady, the more uniform of distribution of mobile users is, the larger the information entropy is, and lower the success rate of inference attack is.

Average Size of Anonymous Edge Set. It represents the amount of road sections in the AES, the smaller it is, the smaller the cost of network transition and query cost are. Thus the average size of AES can reflect the communication cost and query cost to some extent.

Average Query Cost. It can be evaluated by average query candidate set and average query time [8]. Average query candidate set is the average values of returned result sets that are received by LPS. Average query time is the average time of LBS server to complete the query.

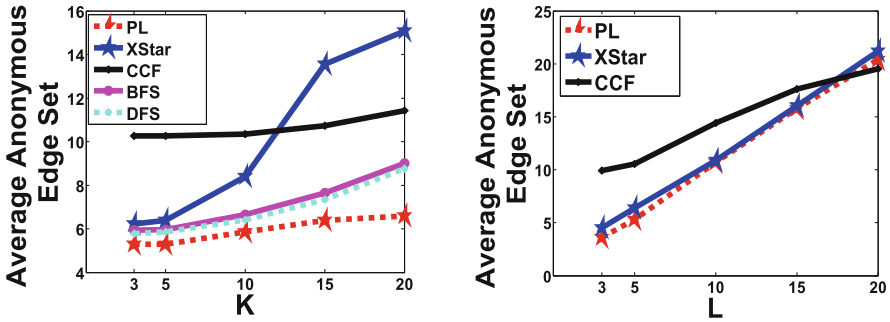
Privacy Success Rate. It is the ratio of successful queries and the total queries [14], and success rate SR = Q_num' /Q_num, where Q_num is the amount of total queries, and Q_num' is the amount of successful queries in personalized location privacy preservation.

5.3 Analysis of Experiment Results

The experiments are mainly evaluated in average information entropy, average size of AES, average query cost, average query time and anonymous success rate for XStar [6], PL, CCF [8], BFS [7] and DFS algorithm [7]. For CCF, we only consider simple road networks; for BFS and DFS, since they don't consider L, we only consider the case when L takes mean value.

From Fig. 3(a) we can see that the average size of AES of PL, BFS and DFS are similar, but XStar and CCF are higher. Since PL uses selection rate, the distribution of users is uniform; for BFS and DFS, they don't have any redundant road sections. For CCF and XStar, they use extra structures, which make the generated AES large.

From Fig. 3(b) we can see that the tendencies are similar. Since the user privacy requirement can be satisfied easily, the influencing factor becomes L, and the tendencies become nearly linear growth. For CCF, it finds many cycles, thus needs more edges when L is small.



(a) L takes mean value (b) K takes mean value

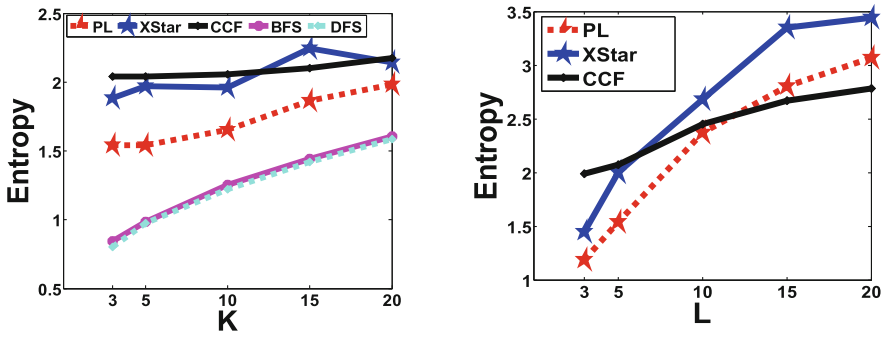
Fig. 3. The relation between average size of AES and privacy parameters

As Fig. 4(a) shows, the average information entropy for XStar and CCF are the best since they have large AESs; PL has small anonymous set, but uses selection rate to balance the distribution in the AES, thus the information entropy is relatively high. For BFS and DFS, they don't consider the balance of distribution and this result in lower information entropy.

From Fig. 4(b) we can see average information entropy is increasing gradually and all of them are similar, since AESs are similar and have uniform distributions when K is steady.

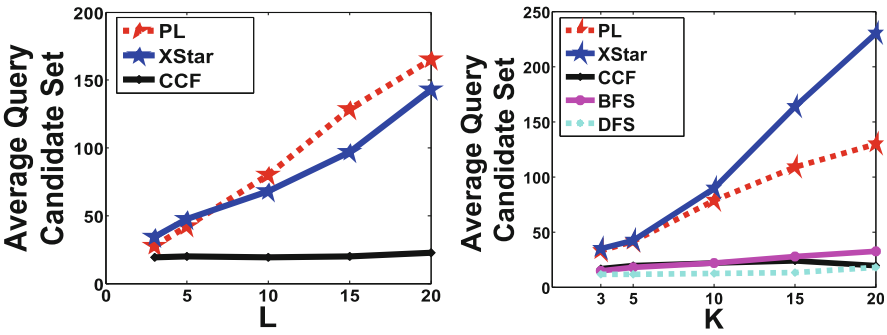
From Fig. 5 we can see that CCF is best for two conditions of average query candidate set, and PL has larger candidate set than CCF, but smaller than XStar in some cases. Since BFS and DFS have smaller AESs, the average size of query candidate set is also small.

From Figs. 5 and 6, we can see the candidate set of CCF is small, but the average query time is long, since the cycles in it are large and it has few open nodes in AES. The average candidate set and average query cost of PL are larger than XStar and CCF



(a) L takes mean value (b) K takes mean value

Fig. 4. Relation between average information entropy and privacy parameters

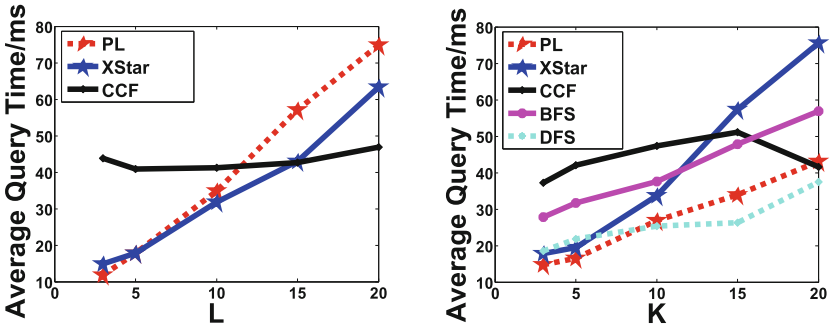


(a) K takes mean value (b) L takes mean value

Fig. 5. The relation between average query cost and K or L

when K takes mean value, because the connectivity of its AES becomes smaller and the amount of open nodes becomes larger with L increasing; when L takes mean value, PL is better since connectivity is steady and AES is smaller; for XStar, it has large AES which leads to large query candidate set and query time; for BFS and DFS, the small AESs and less open nodes lead to small average query candidate set, and also they have less query time since they have less open nodes.

From Fig. 7 we can see that with the increase of K, the success rate of XStar is apparently decreasing, since generated stars can bring redundant sections. The success rate of PL algorithm is steady and high since it considers both K-anonymity and L-diversity and avoids the shortage of redundant road sections; the success rate of PL can keep in 100 % since when L is steady and K is not large, we can always find satisfied AES, which means the query is successful. For CCF algorithm, the generated cycles have lots of road sections, which decreases success rate; for BFS and DFS algorithms, when the value of K is small, the amount of edges is smaller than L, thus they don't satisfy privacy degree,



(a) K takes mean value (b) L takes mean value

Fig. 6. The relation between average query time and K or L

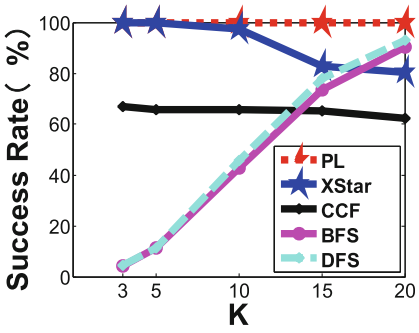


Fig. 7. The relation between privacy success rate and K

which leads to low success rate. From Fig. 7 we can see that the success rate of PL algorithm is higher than those in CCF, XStar, BFS and DFS.

6 Conclusions

Based on the practical road networks, we propose PL algorithm to implement the personalized location privacy preservation. When PL groups the users according to user's privacy parameters K and L, it considers the selection rate of road sections, which can satisfy K and L easily and lead to lower query cost. Additionally, it makes the distribution of users in each road section more uniform, which decreases the success rate of inference attack and increases the privacy degree. The experiment results show that PL can provide higher quality of service in personalized location privacy preservation than some typical algorithms like XStar and CCF algorithms.

Acknowledgments. This work was partially supported by the Natural Science Foundation of China (No. 61272403), by the Fundamental Research Funds for the Central Universities (No. 10561201474). We also appreciate Yaohui Zheng and Kai Tian for their kindly help of the experimental analysis and programming.

References

1. Mokbel, M.F.: Privacy in location-based services: state-of -the-art and research directions. In: Proceedings of the 8th International Conference on Mobile Data Management, p. 228 (2007)
2. Pedreschi, D., Bonchi, F., Turini, F., Verykios, V.S., Atzori, M., Malin, B., Moelans, B., Saygin, Y.: Privacy protection: regulations and technologies, opportunities and threats. In: Giannotti, F., Pedreschi, D. (eds.) Mobility, Data Mining And Privacy, pp. 101–119. Springer, Berlin (2008)
3. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
4. Gruteser, M., Grunwald, D.: Anonymous usage of location- based services through spatial and temporal cloaking. In: Proceedings of the first International Conference on Mobile Systems, Applications, and Services. San Francisco, CA, USA, pp. 163–168 (2003)
5. Ardagna, C.A., Cremonini, M., Damiani, E., De Capitani di Vimercati, S., Samarati, P.: Location privacy protection through obfuscation-based techniques. In: Barker, S., Ahn, G.-J. (eds.) Data and Applications Security 2007. LNCS, vol. 4602, pp. 47–60. Springer, Heidelberg (2007)
6. Wang, T., Liu, L.: Privacy-aware mobile services over road networks. Proc. VLDB Endow. **2**(1), 1042–1053 (2009)
7. Mouratidis, K., Yiu, M.L.: Anonymous query processing in road networks. IEEE Trans. Knowl. Data Eng. TKDE **22**(1), 2–15 (2010)
8. Xue, J., Liu, X., Yang, X., et al.: A location privacy preserving approach on road networks. Chin. J. Comput. **34**(5), 865–878 (2011). (In Chinese)
9. Mokbel, M.F., Chow, C.Y., Aref, W.G.: The new casper: query processing for location services without compromising privacy. In: Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, pp. 763–774 (2006)
10. Hong-Yun, X., Jun, X., Gong, Y.-J., Meng-Zhen, X.: Algorithms to generate location privacy area based on location privacy protection with spatial cloaking. J. S. Chin. Univ. Technol. (Nat. Sci. Ed.) **42**(1), 97–103 (2014). (in Chinese)
11. Machanavajjhala, A., Gehrke, J., et al.: L-diversity: privacy beyond k-anonymity. In: Proceedings of 22nd International Conference on Data Engineering, Atlanta, Georgia, USA, pp. 24–36 (2006)
12. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**(6), 1554–1563 (1966)
13. Brinkhoff, T.: A framework for generating network-based moving objects. GeoInfomatica **6** (2), 153–180 (2002)
14. Pan, X., Xiao, Z., Meng, X.: Survey of location privacy-preserving. J. Front. Comput. Sci. Technol. **1**(3), 268–281 (2007). (In Chinese)