Location Privacy-Preserving Method Based on Degree of Semantic Distribution Similarity

Rui Liu[1,2], Kaizhong Zuo[1,2(✉)], Yonglu Wang[1,2], and Jun Zhao[1,2]

[1] College of Computer and Information, Anhui Normal University, Wuhu 241002, Anhui, China
zuokz@ahnu.edu.cn

[2] Anhui Provincial Key Laboratory of Network and Information Security, Anhui Normal University, Wuhu 241002, Anhui, China

Abstract. While enjoying the convenience brought by location-based services, mobile users also face the risk of leakage of location privacy. Therefore, it is necessary to protect location privacy. Most existing privacy-preserving methods are based on *K*-anonymous and *L*-segment diversity to construct an anonymous set, but lack consideration of the distribution of semantic location on the road segments. Thus, the number of various semantic location types in the anonymous set varies greatly, which leads to semantic inference attack and privacy disclosure. To solve this problem, a privacy-preserving method is proposed based on degree of semantic distribution similarity on the road segment, ensuring the privacy of the anonymous set. Finally, the feasibility and effectiveness of the method are proved by extensive experiments evaluations based on dataset of real road network.

Keywords: Location-based services · Road network · Semantic location · Privacy-preserving

1 Introduction

With the rapid development of communication technology and mobile positioning technology, users can use mobile devices such as in-vehicle terminals and mobile phones to obtain their locations anytime and anywhere, thereby application based on Location-based Services (LBS) have become more and more widespread [1–3]. If users wish to receive information from LBS, they have to share their exact location. For example, how to go to the nearest hospital? Meanwhile, users face the risk of leakage of location privacy [4]. More sensitive personal information can be stolen by the attacker. Therefore, how to solve the problem of leakage of location privacy in LBS has attracted the attention of scholars at home and abroad.

Currently, several schemes have been proposed by scholars [5–9] to protect the location privacy of users. For example, *K*-anonymous algorithm is usually used in Euclidean space [5, 6], where users can move freely. These algorithms construct anonymous set including *k* users instead of the exact location of the user, which makes it difficult for an attacker to distinguish the exact user from other anonymous users. However, the security of the *K*-anonymous algorithm is compromised when attackers

J. Zeng et al. (Eds.): ICPCSEE 2020, CCIS 1257, pp. 118–129, 2020.
https://doi.org/10.1007/978-981-15-7981-3_9

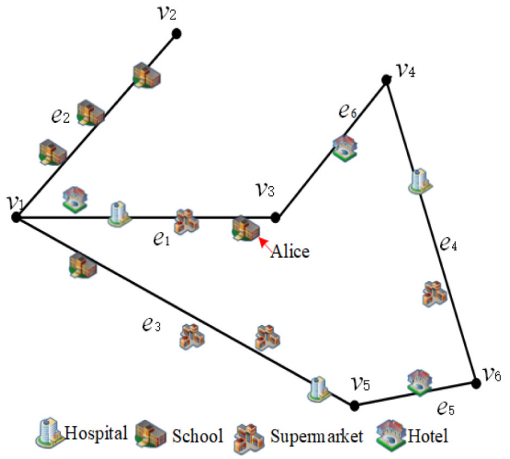
mined road network information and semantic location information. Therefore, it is extremely important to propose location privacy-preserving scheme on the road network.

Location privacy-preserving schemes on the road network are mostly based on (K, L)-model, which means the anonymous set not only includes K anonymous users, but also includes L road segments. Chow et al. [10] designed a location privacy-preserving method on road network, which blurs the user's exact position into several adjacent road segments, and considers the query cost and query quality when constructing anonymous set. Pan et al. [11] constructed a unified set of anonymous users for users, and formed anonymous region through connected road segments. However, these methods do not take the information of semantic location on road network environment into consideration. Li et al. [8] used road network intersections and semantic locations to divide the road network into independent non-overlapping network Voronoi units, and optionally added neighboring Voronoi units to form anonymous region to meet semantic security. Xu et al. [12] proposed a global and local optimal incremental query method based on location semantic sensitivity rate. By reducing the number of unnecessary locations, the query overhead is reduced, and the user service quality is improved.

In the above methods guarantee some degree of privacy. Nevertheless, they all share a common drawback. The semantic location is represented by the nearest road network intersection, which is given semantic location information (semantic location type, popularity, sensitivity, etc.). If there are many different types of semantic locations in the intersection of the road network, it will be caused the problem that semantic locations cannot be represented in the road network environment. Chen et al. [13] proposed a privacy-preserving method based on location semantics, which fully considers the user's personalized privacy requirements. In this method, semantic location is directly distributed on the road segment, which makes the road network model more realistic. But the distribution of semantic location types on the road network is not considered, so the number of various semantic location types in the anonymous set varies greatly, which leads to semantic inference attack illustrated by the following example.

Example 1. Figure 1 shows the scenario that a user named Alice requests services through a mobile phone with GPS on road e_1. In order to prevent Alice's location leakage, the method based on K-anonymous and L-segment diversity protects Alice's location privacy. In our example, we assume $K = 45$, $L = 3$, then the anonymous set may be e_1, e_2, e_3. Unfortunately, it is easy for an attacker to infer that Alice is at school, and knows that Alice's identity information may be a student or a teacher. Because the user's semantic location type school accounts for a large proportion of anonymous sets. Therefore, even though Alice's location blur into several adjacent road segments, it's easy to a semantic inference attack.

The rest of the paper is organized as following. We introduce some necessary preliminaries such as fundamental concepts and models in Sect. 2, whereas Sect. 3 introduces the algorithms. The experimental settings and results of our experiments are illustrated in Sect. 4. Finally, Sect. 5 concludes this paper.



(a)Simplified Road Network

Edge Id	*Number of users*
e_1	10
e_2	15
e_3	25
e_4	20
e_5	10
e_6	15

(b) Number of users on the edge

Fig. 1. A snapshot of mobile users in road network

2 Preliminaries

2.1 Related Definition

Definition 1 (Semantic location). $loc = (slid, eid, x, y, type)$ is the semantic location on the road network, *slid* denotes the number of the semantic location, *eid* denotes the number of the road segment where the semantic location is located, x and y are the coordinate of the semantic location. And the *type* is the semantic location type, which contains m types and $Type = \{type_1, type_2, \ldots, type_m\}$ is types of the semantic location.

Definition 2 (Semantic location popularity). We use it to describe the popularity of a semantic location type. Every semantic location type $type_i \in Type$ has their own popularity pop_{type_i}. The set $POP = \{pop_{type_1}, pop_{type_2}, \ldots, pop_{type_m}\}$ indicates the popularity of all semantic location type.

Definition 3 (Vector of road segment popularity). It describes all the semantic location information on the road segment, which is consisted of the popularity and number of semantic locations. And it can be denotes as $\overrightarrow{e_i}$.

Definition 4 (Semantic Road network). A road network is an undirected graph $G = (V, E)$ with a node set V and an edge set E. A node $v_i \in V$ denotes the intersection of the road segment on the road network. While an edge $e_i = (eid, v_i, v_j, \overrightarrow{e_i}) \in E$ is the road segment, connects two nodes v_i and v_j, with *eid* is the road segment number, $\overrightarrow{e_i}$ denote the vector of the road segment popularity.

Example 2 (Semantic Road network). Figure 1 (a) displays a simplified semantic road network model, in which an edge is associated with the vector of the road segment popularity. For example, we assume that the popularity of hospitals, schools, supermarkets, and hotels are 0.3, 0.3, 0.3, 0.1. The vector of road segment popularity on the road segment e_1 and e_2 is $\overrightarrow{e_1} = [0.3, 0.3, 0.3, 0.1]$ and $\overrightarrow{e_2} = [0, 0.9, 0, 0]$, respectively.

Definition 5 (Degree of semantic distribution similarity). It describes the degree of semantic distribution similarity between road segments, which is determined by the vector of road segment popularity and in a range of values [0, 1].

$$\delta_{\overrightarrow{e_i},\overrightarrow{e_j}} = \frac{\sum_{k=1}^{n} \left| e_{ik} - e_{jk} \right|}{\sum_{k=1}^{n} e_{ik} + \sum_{k=1}^{n} e_{jk}}, \delta_{\overrightarrow{e_i},\overrightarrow{e_j}} \in [0, 1] \tag{1}$$

Where $\overrightarrow{e_i}$ and $\overrightarrow{e_j}$ instead the vector of road segment popularity of the road segment e_i and e_j, respectively. The smaller $\delta_{\overrightarrow{e_i},\overrightarrow{e_j}}$ is, the higher degree of semantic distribution similarity of road segments is.

Definition 6 (Semantic location sensitivity). It denotes the sensitivity of semantic location type. Every user can set their own sen_{type_i} for each type of semantic location $type_i \in Type$ freely, and $Sen_u = \{sen_{type_1}, sen_{type_2}, \ldots, sen_{type_3}\}$ denotes sensitivity of all semantic location types.

Definition 7 (Anonymous set). It is a cloaked set of several adjacent road segments such that satisfies the user specified privacy requirements.

Definition 8 (Anonymous set popularity). The popularity Pop_{AS} of the anonymous set,

$$Pop_{AS} = \sum_{i=1}^{|Type|} \frac{|AS.locs.type = type_i|}{|AS.Locs|} pop_{type_i} \tag{2}$$

Definition 9 (Anonymous set sensitivity). The sensitivity Sen_{AS} of the anonymous set,

$$Sen_{AS} = \sum_{i=1}^{|Type|} \frac{|AS.locs.type = type_i|}{|AS.Locs|} sen_{type_i} \tag{3}$$

|Type| in above is the number of semantic location types contained in the anonymous set; |AS.Locs| is the number of semantic locations included in the anonymous set.

Definition 10 (Anonymous set privacy). The privacy PM_{AS} of the anonymous set,

$$PM_{AS} = \frac{Pop_{AS}}{Sen_{AS}} \tag{4}$$

Obviously, the popularity and sensitivity of anonymous set directly affect the privacy of anonymous set. The popularity of anonymous set is higher and the sensitivity is lower, the privacy of the anonymous set is higher.

Definition 11 (Privacy requirement). The user's privacy requirements is denote as $PR(UN, SN, \delta, Sen_u)$. *UN and SN* denotes the user-defined lowest number of mobile users and road segments, respectively; with δ is the highest value of degree of semantic distribution similarity on road segments; and Sen_u is sensitivity of different semantic location types.

2.2 System Architecture

Figure 2 shows the classic centralized server architecture [14], which mainly contains three components: user, anonymous server and LBS server. In this architecture, users can obtain their location information from the base station. Then, send it to the anonymous server together with query content and privacy requirements (step ①). Subsequently, the anonymous server uses the semantic location privacy-preserving module to blur the user's location into a set of road segment that meet the user's privacy requirement, and sends anonymous query to the LBS server (step ②). After the LBS server gets the candidate results for anonymous query and sends it to the anonymous server (step ③). Finally, the anonymous server computes the candidate results through the filter module and delivers exact result to the query user (step ④).

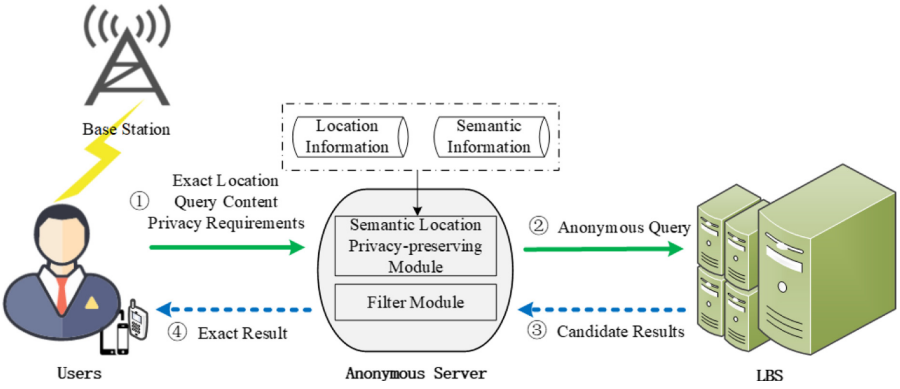


Fig. 2. System architecture

3 Semantic Location Privacy-Preserving Method

The method selects the appropriate adjacent road segments to construct an anonymous set according to user-defined privacy requirements. And it consists of two algorithms.

Algorithm 1 (ORSSA) is used to determine the optimal road segment in this paper. Calculate the popularity vector of the user's current road segment and adjacent road segments set, and take the road segments to candidate road segment sets which less than or equal to δ. Subsequently, select an optimal road segment from these road segment to join the anonymous set. Here, the optimal road segment means that the privacy level of the anonymous set composed of the road segment is highest. Detail process is depicted as following:

Algorithm 1 Optimal Road Segment Selection Algorithm (ORSSA)

Input: user u, current anonymous set CAS, *adjacent road segments set* $AdjacentEdges_{\text{set}}$, threshold δ, sensitivity set $Sens$, *popularity set POP*

Output: $OEdge$

1) $OEdge = \varnothing$; $CandEdges_{set} = \varnothing$; $PM_{set} = \varnothing$;
2) $MAX=0$;
3) for each edge in $AdjacentEdges_{\text{set}}$
4) assign the edge where user u located to e_1, calculate the vector of road segment popularity $\overrightarrow{e_1}$;
5) assign edge to e_2, calculate the vector of road segment popularity $\overrightarrow{e_2}$;
6) if $\sum_{k=1}^{n}|e_{1k} - e_{2k}| / (\sum_{k=1}^{n} e_{1k} + \sum_{k=1}^{n} e_{2k}) \le \delta$ then
7) $CandEdges_{set} = CandEdges_{set} \bigcup edge$;
8) end if
9) end for
10) for each edge in $CandEdges_{set}$
11) $PM = Pop_{(CAS \cup edge)} / Sen_{(CAS \cup edge)}$;
12) $PM_{set} = PM_{set} \cup PM$;
13) $MAX = \{PM \mid \max\{PM \in PM_{set}\}\}$, and the corresponding edge assigned to $OEdge$;
14) end for
15) return $OEdge$

The next algorithm (DSDSPPA) is a privacy-preserving algorithm based on degree of semantic distribution similarity of road segments. The idea of our method is to start from the road segment where the user is located. If this road segment meets the user's privacy requirements, we return it as anonymous set to the LBS server. Otherwise, we pick the optimal road segment from all adjacent road segment until the user's privacy requirements are met. Detail process is depicted as following:

Algorithm 2 Privacy-preserving algorithm based on degree of semantic distribution similarity (DSDSPPA)

Input: user u, privacy requirements PR, popularity set POP
Output: Anonymous set AS
1) AS = ∅ ;
2) assign the edge where user u located to $OEdge$;
3) $AS = AS \cup OEdge$;
4) while $|AS.users| < PR.UN$ or $|AS.edges| < PR.SN$
5) $AdjacentEdges_{set} = GetEdges(AS)$;//find adjacent road segments of AS
6) $OEdge=ORSSA(AS, AdjacentEdges_{set}, PR.\delta, PR.Sen_u, POP)$;
7) $AS = AS \cup OEdge$;
8) Adjacent$Edges_{set} = \varnothing$;
9) *end while*
10) return *AS;*

4 Experimental Evaluation

Our algorithms are executed in java based on MyEclipse environment and all experiments are performed on Intel (R) Core(TM) i5-9400 CPU @ 2.90 GHz and 16 GB main memory with Microsoft Windows 10 Professional.

4.1 Datasets and Parameter

(1) Datasets

In this paper, we use the road network in California which includes 21048 vertices and 21693 edges. And real road network dataset contains semantic location of various categories, e.g., hospital, park, airport, bar, building [15]. The second experimental dataset used in this paper was collected from Gowalla, which has more than 6442890 check_ins made by users over the period of Feb.2009-Oct.2010 [16]. Then, we filter the user's check_ins data in California and calculate the popularity of different semantic location types.

(2) Query generator

We generate 10000 mobile users through by the Network Generator of moving objects [17] and choose 1,000 users send a query randomly. Table 1 depicts the parameters of our experiment.

Table 1. Parameter setting

Parameter	Default values	Range
PR.UN	25	[10,40]
PR.SN	6	[3,15]
$PR.SN_{\max}$	20	
δ	0.7	[0.1,1]
The number of mobile users	10000	
The number of users that issue queries	1000	
The number of semantic location types	63	

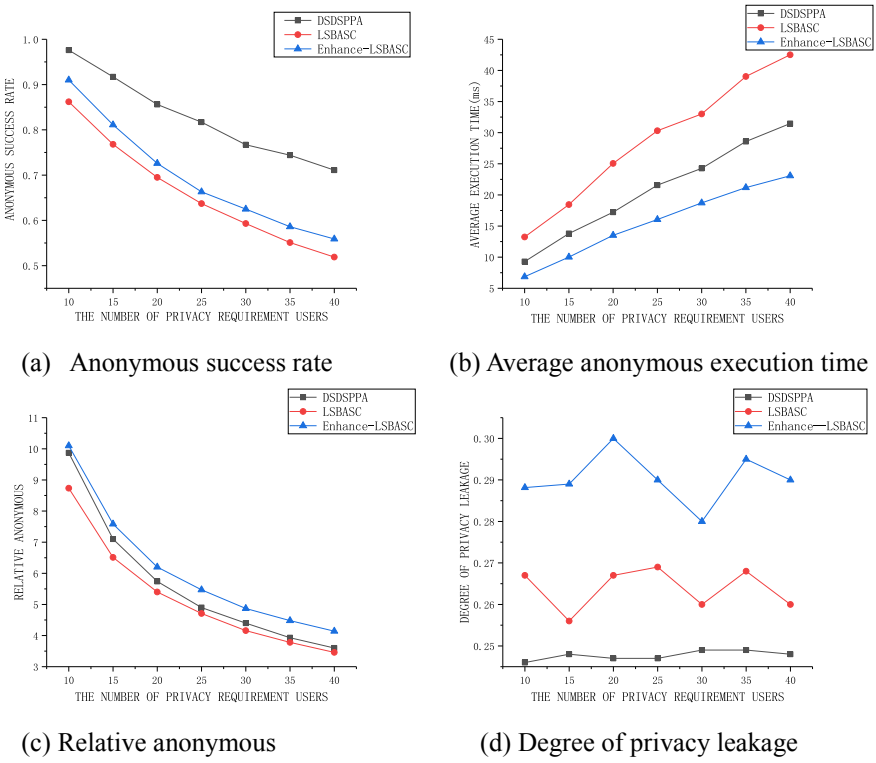
4.2 Experimental Results

Assess the feasibility and effectiveness of the algorithms DSDSPPA, we compare algorithm LSBASC [13] and Enhance-LSBASC [18] from anonymous success rate, average anonymous execution time, relative anonymous and degree of privacy leakage.

(1) Effect of Number of Privacy Requirement Users

Figure 3 shows the effect of different number of privacy requirement users on three algorithms when $PR.SN = 6$, $\delta = 0.7$, $PR.SN_{\max} = 20$. As can be seen from Fig. 3(a) that three algorithm show decreasing trend on the aspect of anonymous success rate. With the increase of number of privacy requirements users, more road segments are needed to anonymous set. So that the number of road segments is higher than $PR.SN_{\max}$ lead to the failure of anonymity. However, the anonymous success rate of DSDSPPA algorithm is always higher than the other algorithms. In Fig. 3(b), the average anonymous execution time of the algorithm DSDSPPA is between the other algorithms, but it is always lower than the algorithm LSBASC. And the average anonymous execution time of all algorithms show increasing trend. The reason of this is that more users must be added in anonymous set in order to meet privacy requirement, so does more road segments. The algorithm DSDSPPA and LSBASC only add one road segment every time, while the algorithm Enhance-LSBASC adds several adjacent road segments, so the average anonymous execution time is less than the algorithm DSDSPPA and LSBASC.

Figure 3(c) shows that the relative anonymity of algorithm DSDSPPA is between the other algorithms. Because the algorithm Enhance-LSBASC chooses several adjacent road segments to the anonymous every time, the number of mobile users contained in the anonymous set is higher than algorithm DSDSPPA and LSBASC. Figure 3(d) is shown that the privacy leakage degree of the three algorithms. Algorithm DSDSPPA is always lower than the algorithm Enhanced-LSBASC and LSBASC, and the fluctuation degree is smallest. In order to satisfy the number of users with privacy requirements, new adjacent road segments need to be added to the anonymous set. Because the algorithm DSDSPPA selects the road segments with high degree of semantic distribution similarity as the candidate road segments, it balances the number of various semantic positions in the anonymous set, so that the attacker cannot infer the semantic position type of the user. However, the algorithm Enhanced-LSBASC and LSBASC only consider the privacy of the anonymous set, so they fluctuate greatly.



(a) Anonymous success rate

(b) Average anonymous execution time

(c) Relative anonymous

(d) Degree of privacy leakage

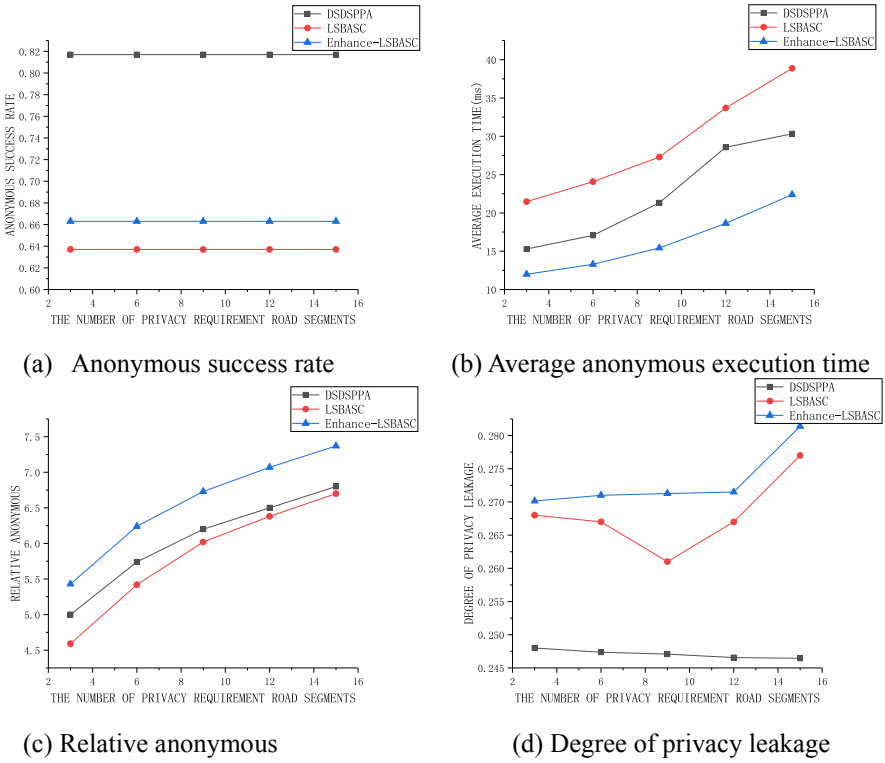
Fig. 3. Results of different number of privacy requirement users

(2) Effect of Number of Privacy Requirement Road Segments

Figure 4 shows the effect of different number of privacy requirement road segments on three algorithms when $PR.UN = 25$, $\delta = 0.7$, $PR.SN_{\max} = 20$. Figure 4(a) shows that different number of privacy requirement road segments don't have an impact on anonymous success rate. In the case of the number of users with privacy requirements remains unchanged, as long as the number of road segments is within the allowed range, the anonymous success rate will not change. According to the experimental result in Fig. 4(b), the average anonymous execution time of all algorithms show increasing trend, but the execution time of DSDSPPA algorithm is between LSBASC and Enhance-LSBASC algorithm and always lower than LSBASC algorithm. In order to meet the number of privacy requirement road segments, more road segments should add to anonymous set. The algorithm DSDSPPA and LSBASC only add one road segment every time, while the algorithm Enhance-LSBASC adds several adjacent road segments, so the average anonymous execution time is less than the DSDSPPA algorithm and LSBASC.

According to the experimental result in Fig. 4(c), the relative anonymity of the three algorithms is increasing. And it can be known that the relative anonymity of the algorithm DSDSPPA is between the other algorithms. With the number of road

segments added to the anonymous set to meet the privacy requirements, at the same time, the number of mobile users is also increasing. Figure 4(d) is shown that the privacy leakage degree of the algorithm DSDSPPA is always lower than the algorithm Enhanced-LSBASC and LSBASC, and the fluctuation degree is smallest. Because the algorithm DSDSPPA selects the road segments with high degree of semantic distribution similarity as the candidate road segments, it balances the number of various semantic positions in the anonymous set.



(a) Anonymous success rate

(b) Average anonymous execution time

(c) Relative anonymous

(d) Degree of privacy leakage

Fig. 4. Results of different number of privacy requirement road segments

5 Conclusion

As constructing anonymous set does not consider the distribution of semantic location types on the road network, the number of various semantic location types in the anonymous set varies greatly, which leads to semantic inference attack and privacy disclosure. Therefore, this paper proposes a location privacy-preserving method based on semantic location information on the road segment. It increases the indistinguishability of users' semantic location types and guarantees the privacy of anonymous sets. Finally, extensive experiments evaluations based on dataset of real road network show our algorithms is effective and feasible.

Acknowledgement. This paper was supported by the National Natural Science Foundation of China under Grant No. 61672039 and 61370050; and the Key Program of Universities Natural Science Research of the Anhui Provincial Department of Education under Grant No. KJ2019A1164.

References

1. Wan, S., Li, F.H., Niu, B., et al.: Research progress of location privacy protection technology. Chin. J. Commun. **37**(12), 124–141 (2016)
2. Zhang, X.J., Gui, X.L., Wu, Z.D.: Review of research on privacy protection of location services. Chin. J. Softw. **26**(9), 2373–2395 (2015)
3. Sun, Y., Chen, M., Hu, L., et al.: ASA: against statistical attacks for privacy-aware users in Location Based Service. Future Gener. Comput. Syst. **70**(70), 48–58 (2017)
4. Zhang, Y., Szabo, C., Sheng, Q.Z.: SNAF: observation filtering and location inference for event monitoring on twitter. World Wide Web **21**(2), 311–343 (2018)
5. Feng, Y., Xu, L., Bo, S.: (k, R, r)-anonymity: a light-weight and personalized location protection model for LBS query. In: ACM Turing Celebration Conference-China, pp. 1–7 (2017)
6. Ma, M., Du, Y.: USLD: a new approach for preserving location privacy in LBS. In: Workshop on Information Security Applications, pp. 181–186 (2017)
7. Cui, N., Yang, X., Wang, B.: A novel spatial cloaking scheme using hierarchical hilbert curve for Location-Based Services. In: Cui, B., Zhang, N., Xu, J., Lian, X., Liu, D. (eds.) WAIM 2016. LNCS, vol. 9659, pp. 15–27. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39958-4_2
8. Li, M., Qin, Z., Wang, C., et al.: Sensitive semantics-aware personality cloaking on road-network environment. Int. J. Secur. Appl. **8**(1), 133–146 (2014)
9. Li, H., Zhu, H., Du, S., et al.: Privacy leakage of location sharing in mobile social networks: attacks and defense. IEEE Trans. Dependable Secure Comput. **15**(4), 646–660 (2016)
10. Chow, C., Mokbel, M.F., Bao, J., et al.: Query-aware location anonymization for road networks. Geoinformatica **15**(3), 571–607 (2011)
11. Pan, X., Chen, W.Z., Sun, Y., et al.: Continuous queries privacy protection algorithm based on spatial-temporal similarity over road networks. Chin. J. Comput. Res. Dev. **54**(9), 2092–2101 (2017)
12. Xu, M., Xu, H., Xu, C.: Personalized semantic location privacy preservation algorithm based on query processing cost optimization. In: Wang, G., Atiquzzaman, M., Yan, Z., Choo, K.-K.R. (eds.) SpaCCS 2017. LNCS, vol. 10656, pp. 153–168. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-72389-1_14
13. Chen, H., Qin, X.: Location-semantic-based location privacy protection for road network. Chin. J. Commun. **37**(8), 67–76 (2016)
14. Wang, Y., Zuo, K., Liu, R., Guo, L.: Semantic location privacy protection based on privacy preference for road network. In: Vaidya, J., Zhang, X., Li, J. (eds.) CSS 2019. LNCS, vol. 11983, pp. 330–342. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37352-8_30
15. Li, F., Cheng, D., Hadjieleftheriou, M., Kollios, G., Teng, S.-H.: On trip planning queries in spatial databases. In: Bauzer Medeiros, C., Egenhofer, M.J., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 273–290. Springer, Heidelberg (2005). https://doi.org/10.1007/11535331_16

16. Cho, E., Myers, S.A., Leskovec, J., et al.: Friendship and mobility: user movement in location-based social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1082–1090 (2011)
17. Brinkhoff, T.: A framework for generating network-based moving objects. GeoInformatica **6** (2), 153–180 (2002)
18. Lv, X., Shi, H., Wang, A., et al.: Semantic-based customizable location privacy protection scheme. In: International Symposium on Distributed Computing and Applications for Business Engineering and Science, pp. 148–154. IEEE Computer Society (2018)