A *k*-anonymous Location Privacy Protection Method of Dummy Based on Geographical Semantics

Yong-Bing Zhang[1,2], Qiu-Yu Zhang[1], Zong-Yi Li[2], Yan Yan[1], and Mo-Yi Zhang[1]
(Corresponding author: Qiu-Yu Zhang)

School of Computer and Communication, Lanzhou University of Technology[1]
No. 287, Lan-Gong-Ping Road, Lanzhou 730050, China
(Email: zhangqylz@163.com)
Gansu Institute of Mechanical & Electrical Engineering[2]
No. 107, Chi-Yu Road, Tianshui, Gansu 741001, China
(Received Aug. 2, 2018; Revised and Accepted Feb. 7, 2019; First Online June 17, 2019)

Abstract

Dummy is one of the main methods used to protect location privacy. In existing methods, the efficiency of dummy generation is low, and the geographical semantic information of location is not fully taken into account. In order to solve these problems, a *k*-anonymous location privacy protection method of dummy based on geographical semantics was proposed in this paper. Firstly, the location data set in the rectangle region containing the real location is obtained from WiFi APs. Secondly, adopting the multicenter clustering algorithm based on max-min distance, some locations are selected. Its geographical distance between them is the farthest, and a candidate set of dummies is generated. Finally, by calculating the edit-distance between geographic name's information of locations, the semantic similarity between any two locations in the candidate set is obtained, and *k*-1 locations with the minimum semantic similarity are selected as dummies. Experimental results show that the proposed method can ensure the physical dispersion and semantic diversity of locations, as well as the improvement of the efficiency of dummy generation. Meanwhile, the balance between privacy protection security and query service quality is achieved.

Keywords: Clustering Center; Dummy; Geographic Semantics; k-*anonymous; Location Privacy Protection; Semantic Similarity*

1 Introduction

With the development of mobile location technology and wireless communication technology, a large number of mobile devices in the market have capacity of GPS precise positioning, which makes location-based service (LBS) become one of the most promising services to mobile users [33]. However, when LBS provide convenience and great benefits to the society, its problem of sensitive information leakage has attached more attentions by many people. Because user's location is shared among different location service providers (LSPs), untrustworthy third parties can easily steal user's privacy via analyzing and comparing these locations' information [35]. For example, through capturing recent users' trace, some information can be analyzed by adversary such as home addresses, workplaces, and health conditions, etc. Therefore, it is necessary to ensure the safety of users' location privacy.

Currently, in order to prevent the leakage of privacy information, many different methods are proposed by experts and scholars, including fuzzy method, encryption method and strategy-based method. Because of the better reliability, the fuzzy method is the most commonly used in the field of location privacy protection, which is mainly realized by means of spatial anonymity or dummy technology. The spatial anonymous method usually needs the help of Fully-Trusted Third Party (TTP) [16]. When a location query service is needed, the mobile user first sends the query request to the TTP, a *k*-anonymous region containing the user's location is generated by the TTP and then it will be sent to the LBS server for query. In this method, if the area of *k*-anonymous region is too large, it not only consumes more time, but also reduces the accuracy of the query result. Meanwhile, TTP is easy to become a bottleneck of system. However, in the dummy-based location privacy protection, TTP and anonymous region are not required, and the dummy locations are generated by mobile clients. Thus, it can compensate the above disadvantages of spatial anonymous methods well.

In the dummy-based location privacy protection, in or-

der to improve the efficiency of dummy location generation and the query service quality, a k-anonymous location privacy protection method of dummy based on geographical semantics was proposed. In this paper, we give full consideration to the geographical semantic information features of locations. Firstly, adopting multi-center clustering algorithm based on max-min distance (MCAMD) [24], a number of cluster centers are generated by clustering calculation, which constitute a candidate set of dummies. Then, using edit-distance [37] to calculate semantic similarity of geographic name's information among elements in candidate set, and k-1 locations with the minimum semantic similarity are selected as dummies. The proposed method can meet semantic l-diversity and physical dispersion of locations, and improve the efficiency of dummies generation. Furthermore, it improves the query service quality.

Our main contributions can be summarized as follows:

1) A dummy selection method considering the geographical semantic information characteristics of locations is proposed, which balanced the contradiction between privacy protection and query quality.

2) A multi-center clustering algorithm based on the max-min distance method is used to generate candidate set of dummies, which can ensure the physical dispersion of the dummies.

3) We calculate the semantic similarity between geographic name's information of locations, and the locations with the smallest semantic similarity is selected as the dummies, which ensures the semantic diversity of the dummies.

The remaining part of this paper is organized as follows. Section 2 reviews related work of location privacy protection. Section 3 gives system model of this paper. Section 4 describes two algorithms and analysis. Section 5 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 6.

2 Related Work

The location privacy protection method is divided into two main categories according to the system architecture, including distributed structure [21]based on Peer-to-Peer (P2P) [23]and central server structure based on TTP [29]. In the distributed structure, location privacy protection is accomplished through collaboration between users. Chow *et al.* [4, 6, 7] proposed a P2P-based spatial anonymity method. In these methods, the k-anonymous privacy protection based on distributed architecture is achieved by using location information of neighbors' node, but the security of the neighbors' node is ignored. The P2P-based scheme is simple and flexible, but which greatly increases various overhead of the smart phone. Furthermore, users are mobile rather than static [34]. In a centralized structure based on TTP, a method of location privacy protection based on TTP is proposed by Zhou *et al.* [36]. This structure mode has good effect of privacy protection, but TTP also needs to be protected. Li *et al.* [12] proposed a location privacy protection scheme based on efficient information cache, which reduces the number of times that the users' access to TTP, the query efficiency is improved, and the probability of information leakage is reduced, but the burden of the mobile client is increased.

In addition, Cheng *et al.* [3] put forward an independent structure model, and users protect location privacy according to their own abilities and knowledge. The structure of this method is simple, which is easy to merge with other structures, but it requires high performance for mobile clients. Li *et al.* [11] put forward a multi-server architecture, users can be divided into different subsets according to the security requirements, and each location server can only obtain partial subset. The concealment of location is improved in this method, but it is mainly suitable for the social network. Mouratidis *et al.* [15] put forward a location privacy protection method based on privacy information retrieval, and its location privacy protection is implemented by using retrieval and encryption. The location privacy is well protected in this method, but it increases the overhead of communication and hardware, and reduces the query service quality. With the maturity and popularity of cloud service technology, Kim *et al.* [10] proposed a location privacy protection method based on searchable encryption. By accessing to the cloud server in the encrypted state, the security of location data and query records is guaranteed, but query efficiency and query accuracy need to be improved further.

In the recent researches, k-anonymity [25] is still the mainstream method of location privacy protection, which was born in the relational database, and its key attribute is dealt with using generalization and fuzzy technology. So none of the records can be distinguished from other k-1 records, and the location anonymity is realized. The method of k-anonymity location privacy protection is mainly divided into spatial region anonymity and dummy anonymity. Gruteser *et al.* [5] proposed a k-anonymity location privacy protection method, and its location privacy is protected by constructing k-anonymous region. The region must meet two conditions: 1) The area of the region reaches a certain value; 2) There are k users in the region. Due to the above tow limitations, the effect of location privacy protection is improved, but all users must have the same location anonymity requirement. Bamba *et al.* [1] put forward a method of grid partition, which provide two algorithms: Top-Down Grid Cloaking algorithm and Bottom-Up Grid algorithm, which can be selected according to the users' needs. Xu *et al.* [27] proved that the size of k-anonymous region has a great impact on the accuracy of query results, which provides guidance for the research of anonymous region partition. On this basis, some anonymous region construction methods with various geometric shapes were presented in [20, 26, 30, 31].

However, these methods have two serious shortcomings: First, it must rely on TTP, but TTP is not absolutely secure, and it's easy to become the bottleneck of the system. Second, the size of anonymous region and the accuracy of query results are a pair of contradiction, and the larger the anonymous region, the better the effect of privacy protection, but the accuracy of the query results will be reduced.

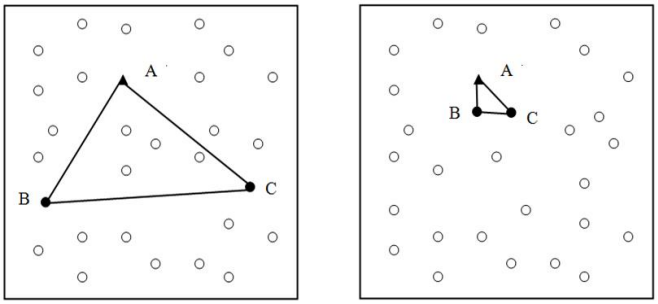
Because of these above serious shortcomings in the spatial anonymity method, the k-anonymity method of dummy has been widely used. The dummy method was first introduced into the location privacy protection by Kido *et al.* [8, 9] in 2005. And then Lu *et al.* [14] proposed a method of randomly adding dummy locations in a circular or rectangular region, users can select dummy locations in the region according to their demands. Several dummy-based privacy protection methods for continuous queries in mobile trajectories were proposed in document [13,28,32]. Niu *et al.* [18] took into account the adversaries' attack with background information, and DLS algorithm and improved DLS algorithm were proposed. Then, Niu *et al.* [19] introduced cache mechanism into the location privacy protection, a cache-based dummy selection algorithm (CaDSA) was proposed, which has improved the query efficiency. Next, Niu *et al.* [17] proposed a mobile location privacy protection scheme named DUMMY-T, which aims to protect user's location privacy from background attacks, the dummy is generated by the dummy location generation (DLG) algorithm, and dummy path is generated by the dummy path construction (DPC) algorithm, which ensure the security of location privacy. Sun *et al.* [22] selected dummy locations by probability estimation, which can prevent probability attack, and solve the problem that attackers can judge the real location information by analyzing historical records.

3 System Model

3.1 Attack Model

In location-based services, common attacks include background attack, probability attack and semantic attack. For background attacks, location privacy is usually protected by eliminating the link between background information and the user's current location. In general, in order to overcome the probability attack, after obtaining the history query record of the query user, the locations with high query probability is used as the dummies to confuse the attacker. However, there are many forms of semantic attack, so the difficulty of protection is large.

In existing study, most methods select dummy according to query probability, which seldom consider the location's geographical semantic information, and adversaries can easily obtain user' location by analyzing the geographical semantic information. As shown in Figure 1, solid triangle A represents real position, Hollow circle represents dummy candidate set, and solid circle B and C represent the selected dummy locations.



(a) (b)

Figure 1: A sample of the location similarity attack: (a) Semantic features; (b) Geographical features

As shown in Figure 1(a), A, B and C are the selected dummy locations, assuming that the three locations are in the hospital, and adversaries can easily identify that users have health problems through semantic analysis. The selected dummies are too close to the real location in Figure 1(b), adversaries can easily find the exact location of the user by computing geographical distance. Therefore, the selection of dummies should consider the geographical semantic information of the location as much as possible, which can ensure the physical dispersion and semantic diversity of all locations including the real location, and further improve the effect of location privacy protection.

In order to prevent location privacy leakage due to geographical semantic attacks, Chen *et al.* [2] proposed a dummy selection method based on semantic-aware. The physical dispersion and semantic diversity of dummies are guaranteed. However, this method needs to repeatedly calculate the physical distance and semantic distance between all locations, and the efficiency is relatively low when location data is large. Furthermore, it needs construct semantic tree for locations in WiFi APs to compute the semantic distance, the burden of WiFi APs and the time of preprocessing is increased, and the service quality is reduced.

3.2 System Structure

In the TTP-based central server model, if users initiate more queries, TTP is easy to become a system bottleneck. Furthermore, TTP is not absolutely safe and reliable. Once TTP is attacked, all locations privacy will be leaked. So, we use a system model without TTP in this paper, and the generation of dummies and the sending of query requests are all accomplished by the mobile client. The system structure model is shown in Figure 2.

In this system structure, the mobile user obtains the location information of the region including the real location from WiFi APs, as shown in Figure 3 and Figure 4. Firstly, by adopting the MCAMD algorithm, a number of cluster centers are generated in mobile client. These locations are the farthest from each other, the dummy can-

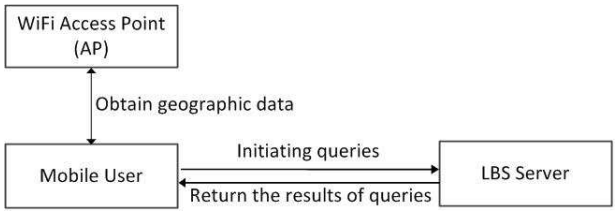


Figure 2: System structure model

didate set is generated from them. Then, the semantic similarity is calculated for the location information of the candidate set, and the k-1 locations with the minimum semantic similarity are selected as the dummies. Finally, the mobile user sends k-1 dummies and real location to the LBS server to query.

Figure 3: Selected location region

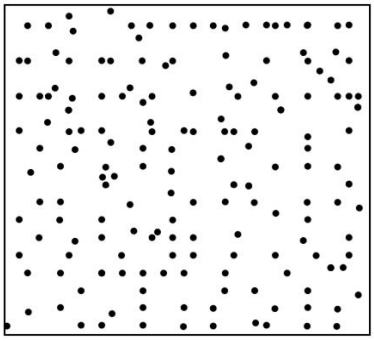


Figure 4: Geographical location in the region

3.3 Definition

Definition 1. *Let R_s represents the selected rectangular area, and $S_n = \{l_1, l_2, \cdots, l_n\}$ represents the set of locations in the rectangle region.*

Definition 2. *Let l_{phi} represents the physical distance between any two locations, and l_{sem} represents the set of locations in the rectangle region.*

Definition 3. *Let $S_1 = \{l_1, l_2, \cdots, l_m\}$ represents the candidate set that satisfies physical dispersion, and $S_2 = \{l_1, l_2, \cdots, l_{k-1}\}$ represents the dummy set that satisfies semantic diversity. Let l_{real} represents the user's location, and the location result set includes the dummy set S_2 and real location l_{real}.*

Definition 4. *If the semantic similarity between l_i and l_j satisfies the following conditions: $1 - \frac{|SEM_{t_i}|}{C_k^2} \geq \theta$, where $SEM_{t_i} = \{l_{sem} | l_{sem}(l_i, l_j) \leq l\}$, k=$|RS_{t_i}|$, C_k^2 is a combination formulas, l is the default threshold of the semantic diversity. Then, the result set RS_{t_i} is called a θ-security set, and the purpose of privacy protection is to get the maximum value of θ by 1, the semantic similarity between l_i and l_j is equal or less than 0.2.*

4 Algorithmic Description

The proposed method of dummy generation is implemented by the following two algorithms: The dummy candidate set S_1 is generated through cluster calculation in Algorithm 1. In Algorithm 2, the dummy set S_2 is generated by calculating semantic similarity of locations in the candidate set S_1.

4.1 Algorithm 1

Algorithm 1: Calculating physical distance and obtaining dummy locations set.

Input: Location data set S_n, demand parameter m.

Output: Generate a dummy candidate set S_1.

Step 1: Given γ value, $0 < \gamma < 1$.

Step 2: The real location l_{real} is taken as the first cluster center Z_1.

Step 3: Find the location that it is the farthest location from Z_1, which is treated as a second Cluster center Z_2.

Step 4: For each l_i of the remaining objects in S_n, its distance to Z_1 and Z_2 is D_{i1} and D_{i2}. Assumed that D_{12} is the distance between Z_1 and Z_2, if $D_i = max\{min(D_{i1}, D_{i2})\}(i = 1, 2, \cdots, n)$ and $D_i > \gamma \cdot D_{12}$, then take l_i as the third Cluster center Z_3.

Step 5: And so on, get all the v clustering centers that conforms to the conditions. When max-min distance is lower than $\gamma \cdot D_{12}$, the calculation for finding the cluster center is finished.

Step 6: Suppose v represents the number of cluster centers obtained by calculation, judge:

1) If $v \geq m$, the algorithm is over, then Step 7,
2) If $v < m$, re-select the γ value, and turn to Step 1.

Step 7: The dummy candidate set S_1 is generated.

4.2 Algorithm 2

Algorithm 2: Calculating semantic similarity and obtaining dummy location result set.

Input: Location candidate set S_1, semantic diversity parameter threshold l.

Output: Location result set S_2.

Step 1: Matching each character of the place name information in turn, and ignore the same prefix characters with the same matching values. Then, get two new character strings A and B.

Step 2: Suppose that the string A contains i characters and it is represented as A=$a_1a_2a_3La_i$; the string B contains j characters and it is represented as B=$b_1b_2b_3Lb_j$.

Step 3: A dynamic programming matrix of i+1 columns and j+1 rows is constructed. The last element obtained from $D[i, j]$ is $ed(A,B)$.

Step 4: If j=0, return i and then exit; if i=0, return j and then exit.

Step 5: The first row is initialized to 0,1,L,i; the first column is initialized to 0,1,L,j.

Step 6: Assign values for each element in the matrix: if a_i=b_i, then $D[i, j]$=$D[i$-1, j-1]; if a_i≠b_i, then $D[i, j]$=1+min($D[i$-1, j-1], $D[i$-1, $j]$, $D[i, j$-1]).

Step 7: Repeat step 6, until all the values in the matrix are obtained, the final edit-distance is $D[i, j]$.

Step 8: Calculating similarity matching index $S(A, B)$ through $D[i, j]$, that is Semantic similarity.

Step 9: Select the k-1 locations with the minimum semantic similarity, and dummy result set S_2 is generated.

4.3 Algorithm 1 Description

Using the MCAMD algorithm to compute cluster center for location geographic coordinates in a square region, several cluster centers are obtained, which are selected as dummy candidate set. The MCAMD algorithm is a clustering algorithm based on heuristic, which takes as far away objects as cluster centers according to Euclidean distance. Firstly, a sample object is used as the first cluster center, and then a sample which is the farthest from the first cluster center is selected as the second cluster center. Then determine the other cluster centers, until there is not new cluster center. After determining all the clustering centers, the clustering sample set including m samples is taken as dummy location candidate set.

The example of cluster center calculation is shown in Figure 5, there are ten locations in the region. According to Algorithm 1, l_1 is selected as first cluster center, and then l_5 is selected as second cluster center, and then third cluster center l_9 is determined. After clustering calculation, three clustering centers are obtained, and the dummy location candidate set is generated.

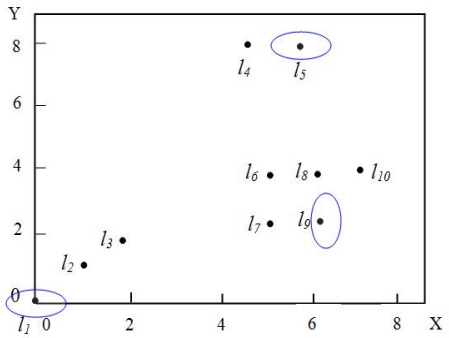


Figure 5: Example of MCAMDA algorithm

When determining the cluster center, the real location is used as the initial cluster center Z_1. If l_i is selected as the i-th clustering center, the conditions must be satisfied.

$$D_i > \gamma \cdot D_{12}(i = 1, 2, \cdots, n) \tag{1}$$

where $D_i = max\{min(D_{i1}, D_{i2})\}(i = 1, 2, \cdots, n)$, $D_{12} = |Z_2 - Z_1|$, γ is the test parameter in the algorithm, it's usual value is $0.5 \leq \gamma < 1$.

4.4 Algorithm 2 Description

Algorithm 2 computes semantic similarity for location information of dummy candidate set. Firstly, according to the characteristics of Chinese geographical names, the same prefix in place name information is eliminated. Then, by calculating semantic similarity for the remaining string of place name through the edit-distance, the efficiency and the accuracy of calculation is improved. For example, "Guangzhou second middle school" and "Guangzhou Tie Yi middle school" are two strings of Chinese place name. The characters of "Guangzhou" do not have any meaning for the calculation of semantic similarity in the two place name strings, and which also affect the accuracy of the calculation results. So "Guangzhou" is eliminated as a prefix in the calculation.

$D[i, j]$ is the edit-distance of the dynamic programming matrix, and the cost of the each edit operation is between 0 and 1. And it can be set different values according to the requirements. In this paper, the value is set to 0 or 1. if $a_i = b_i$, the cost of replacement is 0. Otherwise, the cost of all edit operations is 1. In Equation (2), D is a dynamic programming matrix, which represents the edit-distance between string A= "Second middle school" and string B= "Tie Yi middle school".

$$D = \begin{vmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 1 & 2 & 3 & 4 \\ 2 & 2 & 2 & 3 & 4 \\ 3 & 3 & 3 & 2 & 3 \\ 4 & 4 & 4 & 3 & 2 \end{vmatrix} \tag{2}$$

The edit-distance between the two strings is obtained by calculating, which is $D[i, j]$=$D[4, 4]$=2. Using Equation (3) to calculate the similarity matching index between the strings, that is semantic similarity. The semantic similarity is 0.5.

$$S(A, B) = 1 - \frac{D[i, j]}{max\{|A|, |B|\}} \quad (3)$$

Where $|A|$ and $|B|$ represents the length of two strings respectively, and the maximum length of string S is used to calculate semantic similarity.

At last, according to the Equation (4), the k locations with the minimum semantic similarity including the real location are obtained.

$$Arg\,min(S(l_i, l_j)) \quad (4)$$

4.5 Algorithm Analysis

In this paper, firstly, adopting the MCAMD Algorithm, to select the $m(m > 2k)$ locations with the maximum distance from each other as dummies, dummy candidate set including the real location is generated. Then, by calculating the semantic similarity of the locations in the candidate set, the k locations including the real location are selected as the result set. So physical dispersion and semantic diversity of k locations are ensured.

In Algorithm 1, the candidate set of dummies is generated by clustering calculation, and the physical dispersion between different locations is guaranteed. The clustering results of the algorithm are related to the selection of parameter γ and the first cluster center Z_1, and the real location is used as the initial cluster center. To make sure that the numbers of samples in the candidate set is enough, the number of cluster centers m satisfy the condition of $m > 2k$. In this paper, the initial parameter value of γ is 0.5.

In Algorithm 2, the semantic similarity of geographic name's information is obtained by calculating the edit-distance. In the calculation, the more similar the character in the two strings are, the smaller the edit-distance is, while the greater the semantic similarity is. When the two strings are exactly the same, the edit-distance is 0, and the semantic similarity is 1.

5 Experimental Results and Analysis

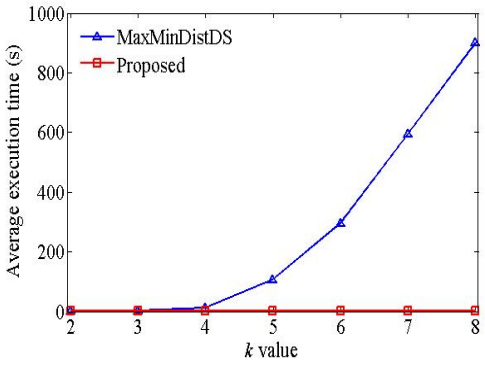
In order to evaluate the performance of the proposed method, we use a real map of Guangzhou from Google maps, and select the 55 WiFi APs in the 8km×8km. The hardware environment of the experiment is as follows: 3.2 GHz Intel Core i5 processor with memory size of 4 GB. The operating system is Windows 7. The proposed method is implemented by Eclipse development platform and Java programming language.

Table 1 is configured for the default parameters of the experiment.

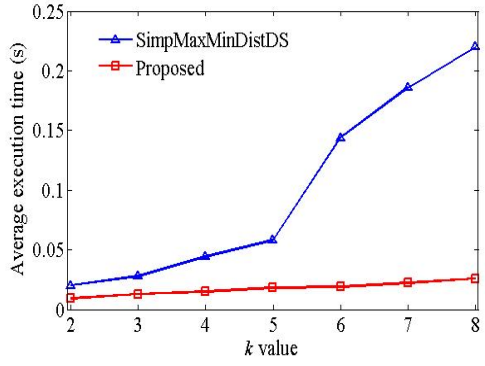
5.1 Average Execution Time

Firstly, the efficiency of the proposed method is verified through experiment. In dummy location selection method considering semantic similarity, we compare the average execution time of dummy locations with MaxMinDistDS [2], SimpMaxMinDistDS [2] and the proposed method. The average execution time of dummy locations of the three methods is shown in Table 2.

In Figure 6, we compare the efficiency of generating dummies with MaxMinDistDS, SimpMaxMinDistDS and the proposed method. As the Figure 6(a) shown, with the increase of k, the MaxMinDistDS algorithm takes much more time than the proposed method. As the Figure 6(b) shown, when $k < 5$, the average execution time of SimpMaxMinDistDS algorithm is slightly larger than that of the proposed method, when $k \geq 5$, the average execution time of SimpMaxMinDistDS algorithm is much larger than that of the proposed method. As can be seen from Figure 6, with the increase of k, the efficiency of the proposed method is more and more advantageous than the other two algorithms.



(a)



(b)

Figure 6: Average generation time of dummy: (a) Comparison between MaxMinDistDS and the proposed; (b) Comparison between SimpMaxMinDistDS and the proposed

In addition, we compare the efficiency of generating dummies with Random [9], Rotation [32], Footprint [28] and DUMMY-T [17] algorithm, as shown in Figure 7.

As can be seen from Figure 7, with the increase of k,

Table 1: Experimental default parameter configuration

Parameter	Value
k	[2, 16]
l	$\leq$0.2
γ	16km×16km
Location set	10000
Space range (km^2)	8×8
WiFi APs Coverage range(m)	800

Table 2: Average execution time vs. k

k	2	3	4	5	6	7	8
MaxMinDistDS	0.07s	1.69s	13s	106.27s	295.41s	592.91s	899.45s
SimpMaxMinDistDS	0.02s	0.028s	0.044s	0.058s	0.0144s	0.186s	0.22s
Proposed	0.009s	0.013s	0.015s	0.018s	0.019s	0.022s	0.026s

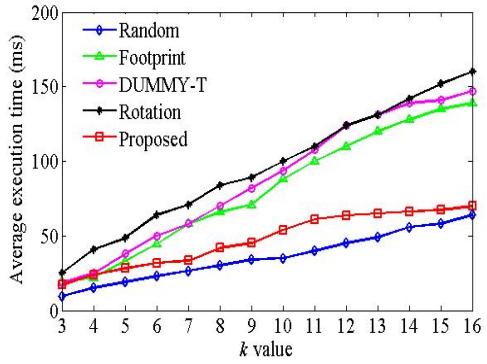


Figure 7: Average generation time of dummy

the average execution time of these algorithms is all increasing. Among them, the average execution time of the proposed and Random algorithm is smaller than other three algorithms. The average execution time of Random algorithm is the least, and the average execution time of the Rotation algorithm is the most. As the Figure 7 shown, when $k\leq4$, the average execution time of Footprint, DUMMY-T and the proposed method are the same. When $k > 4$, The difference in the execution time of the five algorithms is getting bigger and bigger. With the increase of k, average dummy generation time of the proposed method is larger than that of Random algorithm, but it is smaller than the other three algorithms.

Through the analysis of efficiency comparison experiments, it is found that the efficiency of the proposed method is higher than the other three algorithms except Random algorithm. The Random algorithm is randomization, and the effect of privacy protection is relatively poor. The experimental results show that, when the anonymity is large, the proposed method is more efficient under the condition of maintaining good location privacy. Therefore, the efficiency of dummy generation is further improved. And the larger the k is, the better the effect is.

5.2 Comparison of Physical Dispersion

In this paper, we compare the minimum distance of dummies with SimpMaxMinDistDS, MaxMinDistDS and the proposed method. The result is shown in Figure 8.

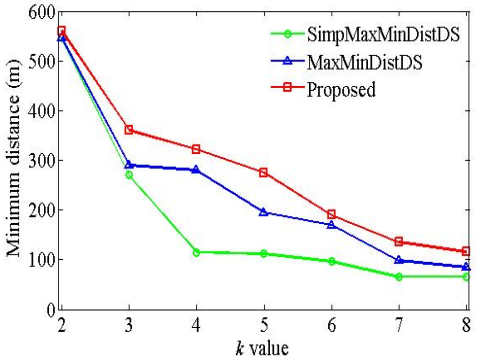


Figure 8: The minimum distance vs. k

As can be seen from Figure 8, the minimum distance between dummies of several methods is mostly reduced with the increase of k, but the minimum distance of the proposed method is obviously larger than the MaxMinDistDS and SimpMaxMinDistDS. Because the proposed method uses the clustering center algorithm in Algorithm 1, and selects the dummies with relatively large distance as the dummy location set, which gives priority to ensuring physical dispersion between locations. However, the MaxMinDistDS first satisfies the semantic diversity and then guarantees the physical dispersion, and the SimpMaxMinDistDS selects the larger in the physical distance and the semantic distance as the dummy location result set. From the experimental result we can see that the proposed method has better physical dispersion.

5.3 Comparison of Semantic Diversity

We compare the semantic diversity with the proposed method, MaxMinDistDS, SimpMaxMinDistDS and DLS [18] through experiment. According to the semantic diversity of the locations in the dummies result set, the θ-secure is obtained, as shown in Figure 9.

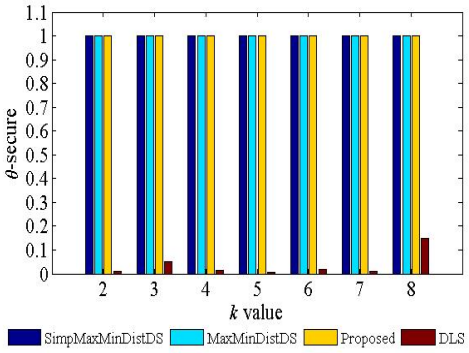


Figure 9: θ-secure vs. k

As can be seen from Figure 9, with the increase of k value, the θ value of algorithm MaxMinDistDS and SimpMaxMinDistDS basically do not change, and always reach 1. The θ value of the proposed method is always the maximum value 1, which can satisfy the requirement of semantic l-diversity. The θ value of DLS is relatively small, and always keeps a lower value. This is because the semantic diversity of geographic location information is considered in MaxMinDistDS, SimpMaxMinDistDS and the proposed method, but in the DLS, the query probability between dummy locations is only considered, and does not consider the semantic information. Moreover, the locations with larger query probability are often in hot areas, and the semantic information between these locations is more similar, thus having greater semantic similarity. Therefore, the semantic diversity of the DLS method is poor, and the θ value is very small.

Through the comparison of experiments, it is found that the proposed method takes less time to generate dummy than other methods. Therefore, the proposed method improves the efficiency of dummy generation, and it further improves the quality of query service. Moreover, through the comparison of experiments, it is found that the physical dispersion and semantic diversity of the dummies selected by this method are better, which can effectively prevent the attack of the opponent who has mastered the characteristics of the geographical semantic information. Therefore, this method not only guarantees location privacy, but also improves the quality of query services, and effectively balanced the contradiction between the effect of location privacy protection and the quality of query service.

5.4 Safety Analysis

In dummy privacy protection, if k dummies are located in one or some areas of concentration, the real location can be easily obtained by reducing the search range. In this case, k-anonymity only meets the requirement in quantity, but does not achieve the effect of anonymity. In the proposed method, the dummies are generated by Algorithm 1, which are distributed uniformly in the region. Therefore, the probability that any location can be distinguished from other k-1 locations is $1/k$, and the effect of anonymity is satisfied. On the basis of geographical distribution, the better the physical dispersion between locations, the better the anonymity.

In semantic attacks, an adversary easily deduces the privacy information of the query user according to the analysis of the semantic relationship between dummies. The greater the difference of the semantic information of geographic name, the better the diversification of location semantics. In Algorithm 2, k locations with the minimum semantic similarity as dummies, it satisfies the requirement of geographic semantics l-diversification.

In conclusion, the proposed method meets the requirements of k-anonymity and l-diversity, and can effectively protect location privacy.

6 Conclusions

In this paper, the issues about location privacy protection based on dummies are discussed, and a k-anonymous privacy protection method of dummy based on geographical semantics was presented. Two algorithms are included in this method: adopting multicenter clustering algorithm based on max-min distance, a number of cluster centers are generated, which constitute a candidate set of dummies in Algorithm 1. In Algorithm 2, by calculating the edit-distance between geographic name's information of locations, the semantic similarity between any two locations in the candidate set is obtained, and dummy location result set with k-1 dummies are generated. We evaluate our algorithms through a series of simulations, which show that our algorithms can ensure the physical dispersion and semantic diversity of locations, protect location privacy effectively, and reduce the time of generating dummy.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61363078, 61762059), the Research Project in Universities of Education Department of Gansu Province of China (No. 2017B-16, 2018A-187), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR)(No. 201700005). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacy grid," in *Proceedings of the 17th International World Wide Web Conference*, pp. 237–246, Jan. 2008.

[2] S. Chen and H. shen, "Semantic-aware dummy selection for location privacy preservation," in *Proceedings of the 16th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 752–759, Aug. 2017.

[3] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar, "Preserving user location privacy in mobile data management infrastructures," *Lecture Notes in Computer Science*, no. 4258, pp. 393–412, 2006.

[4] C. Y. Chow, M. F. Mokbel, and X. Liu, "A peer-to-peer spatial cloaking algorithm for anonymous location-based service," in *Proceedings of the 14th ACM International Symposium on Geographic Information Systems (ACM-GIS'06)*, pp. 171–178, Nov. 2006.

[5] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1th International Conference on Mobile Systems, Applications, and Services*, pp. 31–42, May 2003.

[6] Y. Huang, Z. Huo, and X. F. Meng, "Coprivacy: A collaborative location privacy-preserving method without cloaking region," *Chinese Journal of Computers*, vol. 34, no. 10, pp. 1976–1985, 2011.

[7] R. H. Hwang, Y. L. Hsueh, J. J. Wu, and F. H. huang, "Social hide: A generic distributed framework for location privacy protection," *Journal of Network & Computer Applications*, no. 76, pp. 87–100, 2016.

[8] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proceedings of 1st International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing*, pp. 88–97, Aug. 2005.

[9] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of location privacy using dummies for location-based services," in *Proceedings of the 21st International Conference on Data Engineering Workshops*, pp. 1248, Apr. 2005.

[10] H. I. Kim, H. J. Kim, and J. W. Chang, "A secure knn query processing algorithm using homomorphic encryption on outsourced database," *Data & Knowledge Engineering*, 2017. (https://www.sciencedirect.com/science/article/pii/S0169023X17303476)

[11] J. Li, H. Y. Yan, Z. L. Liu, X. F. Chen, X. Y. Huang, and D. S. Wong, "Location-sharing systems with enhanced privacy in mobile online social networks," *IEEE Systems Journal*, vol. 11, no. 99, pp. 1–10, 2015.

[12] L. Li, J. Hua, S. Wan, H. Zhu, and F. Li, "Achieving efficient location privacy protection based on cache," *Journal on Communications*, vol. 38, no. 6, pp. 148–157, 2017.

[13] H. Liu, X. H. Li, E. M. Wang, and J. F. Ma, "Privacy enhancing method for dummy-based privacy protection with continuous location-based service queries," *Journal on Communications*, vol. 37, no. 7, pp. 140–150, 2016.

[14] H. Lu, C. S. Jensen, and L. Y. Man, "Pad: Privacy-area aware, dummy-based location privacy in mobile services," in *Proceedings of the 7th ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pp. 16–27, Jan. 2008.

[15] K. Mouratidis and M. L. Yiu, "Location-sharing systems with enhanced privacy in mobile online social networks," *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 692–703, 2012.

[16] W. W. Ni, Z. X. Ma, and X. Chen, "Safe region for privacy-preserving continuous nearest neighbor query on road networks," *Journal of Computer Science*, vol. 39, no. 3, pp. 628–642, 2016.

[17] B. Niu, S. Gao, F. H. Li, H. Li, and Z. Q. Lu, "Protection of location privacy in continuous lbss against adversaries with background information," in *Proceedings of the 3rd International Conference on Computing, Networking and Communications (ICNC'16)*, pp. 1–6, Feb. 2016.

[18] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Achieving k-anonymity in privacy-aware location-based services," in *Proceedings of the 33rd Annual IEEE International Conference on Computer Communications (INFOCOM'14)*, pp. 754–762, Apr. 2014.

[19] B. Niu, Q. H. Li, X. Y. Zhu, G. H. Cao, and H. Li, "Enhancing privacy through caching in location-based services," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'15)*, pp. 1017–1025, Apr. 2015.

[20] Z. X. Pei, X. H. Li, H. Liu, and K. Y. Lei, "Anonymizing region construction scheme based on query range in location-based service privacy protection," *Journal on Communications*, vol. 38, no. 9, pp. 125–132, 2017.

[21] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, and J.P. Hubaux, "Hiding in the mobile crowd: Location privacy through collaboration," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 3, pp. 269–279, 2014.

[22] Y. Sun, M. Chen, L. Hu, and Y. Qian, "Asa: Against statistical attacks for privacy-aware users in location based service," *Future Generation Computer Systems*, vol. 70, no. 2017, pp. 48–58, 2016.

[23] E. Troja and S. Bakiras, "Leveraging P2P interactions for efficient location privacy in database-driven dynamic spectrum access," *International Journal of Network Security*, vol. 17, no. 5, pp. 569–579, 2015.

[24] Y. Wu, T. Wang, and J. D. Li, "Clustering parameters selection algorithm based on density for divisional clustering process," *Control and Decision*, vol. 31, no. 1, pp. 21–29, 2016.

[25] M. B. Xie, Q. Qian and S. Ni, "Clustering based k-anonymity algorithm for privacy preservation," *International Journal of Network Security*, vol. 19, no. 6, pp. 1062–1071, 2017.

[26] P. Xie, J. Guo, and Q. Wang, "A-anonymous polygon area construction method and algorithm based on lbs privacy protection," *Journal of Information & Computational Science*, vol. 12, no. 15, pp. 5713–5724, 2015.

[27] J. Xu, X. Tang, H. Hu, and J. Du, "Privacy-conscious location-based queries in mobile environments," *IEEE Transactions on Parallel & Distributed Systems*, vol. 21, no. 3, pp. 313–326, 2010.

[28] T. Xu and Y. Cai, "Exploring historical location data for anonymity preservation in location-based services," in *Proceedings of the 27th Conference on Computer Communications*, pp. 547–555, Apr. 2008.

[29] T. Xu and Y. Cai, "Feeling-based location privacy protection for location-based services," in *Proceedings of the 2009 ACM Conference on Computer and Communications Security*, pp. 348–357, Jan. 2009.

[30] Y. Yang and R. Wang, "Rectangular region k-anonymity location privacy protection based on lbs in augmented reality," *Journal of Nanjing Normal University (Natural science)*, vol. 39, no. 4, pp. 44–49, 2016.

[31] C. Yin, R. Sun, and J. Xi, "Location privacy protection based on improved k-value method in augmented reality on mobile devices," *Mobile Information Systems*, vol. 2017, no. 12, pp. 1–7, 2015.

[32] T. H. You, W. C. Peng, and W. C. Lee, "Protecting moving trajectories with dummies," in *Proceedings of the 9th International Conference on Mobile Data Management*, pp. 278–282, June 2008.

[33] R. Yu, J. Kang, X. Huang, S. Xie, Y. Zhang, and S. Gjessing, "Mixgroup: Accumulative pseudonym exchanging for location privacy enhancement in vehicular social networks," *IEEE Transactions on Dependable & Secure Computing*, vol. 13, no. 1, pp. 93–105, 2016.

[34] H. Zhang, N. Yu, and Y. Wen, "Mobile cloud computing based privacy protection in location-based information survey applications," *Security & Communication Networks*, vol. 8, no. 6, pp. 1006–1025, 2015.

[35] X. J. Zhang, X. L. Gui, and Z. D, "Privacy preservation for location-based services: A survey," *Journal of Software*, vol. 26, no. 9, pp. 2373–2395, 2015.

[36] C. Zhou, C. Ma, and S. Yang, "Location privacy-preserving method for lbs continuous knn query in road networks," *Journal of Computer Research and Development*, vol. 52, no. 11, pp. 2628–2644, 2015.

[37] J. Zhu, B. Hu, and H. Shao, "Trajectory similarity measure based on multiple movement features," *Journal of Wuhan University (Information Science Edition)*, vol. 42, no. 12, pp. 1703–1710, 2017.

Biography

Yong-bing Zhang He is currently a Ph.D. student in Lanzhou University of Technology, and worked at school of Gansu Institute of Mechanical & Electrical Engineering. He received his master degree in electronic and communication engineering from Lanzhou University of Technology, Gansu, China, in 2015. His research interests include network and information security, privacy protection.

Qiu-yu Zhang Researcher/PhD supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is vice dean of Gansu manufacturing information engineering research center, a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, multimedia communication technology.

Zong-yi Li Professor. graduated from Xi'an Jiao Tong University in 1982, and then worked at school of Gansu Institute of Mechanical & Electrical Engineering. He received his master degree in mechanical engineering from Xi'an Jiao Tong University. His research interests include advanced manufacturing technology, engineering CAD technology, expert system, Intelligent manufacturing.

Yan Yan associate professor. received her master degree in communication and information systems from Lanzhou University of Technology, Gansu, China, in 2005. She is currently a Ph.D. student in Lanzhou University of Technology. Her research interests include privacy protection, multimedia information security, uncertain information processing.

Mo-yi Zhang She is currently a Ph.D. student in Lanzhou University of Technology. She received her master degree in communication and information systems from Lanzhou University of Technology, Gansu, China, in 2010. Her research interests include Artificial intelligence, image processing and pattern recognition, robot vision.