Acta Technica *62* (2017), No. 5B, 401–410

Research on semantic method of location privacy protection based on probability

Miaowei Zeng[2], Ming Zhao[2]

Abstract. Mobile users are enjoying LBS (location-based services) under the threat of location privacy at the same time, thus providing effective location privacy protection method is very important.Location privacy protection method based on probability is mainly adopts the space anonymous way.If attackers received anonymous space related background knowledge, especially related to the position of semantic information, will seriously reduce the effectiveness of anonymous.In order to ensure that anonymous area contains plenty of semantic types, and according to the user demand trade-offs between privacy protection and service quality, put forward a kind of semantic method based on probability and the location of the privacy protection.The paper constructs the anonymous area with the maximum entropy and a variety of semantic and meet the demand of customer service, makes the attacker can through the probability or semantic features infer user privacy act.Compared with other algorithms, experimental results show that the proposed location privacy protection based on the probability of semantic method can achieve better location privacy protection effect.

Key words. Location privacy, Query probability, Normalized distance, anonymous.

1. Introduction

With the continuous rapid development of network technology and information technology, Location Based Service (LBS) has developed rapidly and received widespread attention [1,2]. It gives people anywhere access to information and services related to their current location, such as location queries, directions navigation, social entertainment, and more. However, people enjoy the convenience of LBS at the same time, also face the risk of sensitive information disclosure. When the user sends the LBS query, the attacker can infer many realities of the user by analyzing the location information. Such as home address, health status, hobbies and social

[1]Acknowledgment - This paper is supported by National Natural Science Foundation of China with project number: (61073186)(61379057)(61073186)(61309001)(61379110).

[2]Workshop 1 - School of Software, Central South University, ChangSha, 410075, China;email:zhenli45@csu.edu.cn

relationships and more. In recent years, research on privacy protection based on location services has yielded some achievements.Combined with the probability information of anonymous areas, user privacy is protected, but the diversity of semantic information of anonymous areas is ignored. When building anonymous areas, the query probabilities of location units that often have the same semantic information are close to or equal. Therefore, an attacker can infer that the user's real location is in a certain semantic location with greater probability in combination with the semantic information. Second, due to the randomness of the distribution of query probabilities, the formation of anonymous area may cause the anonymous area to be too large, seriously affecting the quality of service. In response to these problems, this paper presents an improved semantic area anonymous building SARB (semantic Anonymouse Region Building) algorithm. The normalized distance is introduced to construct the anonymous area so that the false location in the selected location set is not close to the real location of the user so as to ensure that the anonymous area formed by the false location set contains multiple semantic information. At the same time, by introducing the maximum anonymity interval Amax, the size of the anonymous area to be avoided from being constructed is seriously affected by the size of the anonymous area, so as to better balance the service quality and the privacy protection effect. In this paper, the grid size determined by the level of anonymous h is used as the location unit, and fake location is selected in the entire map range, which can effectively improve the anonymous success rate and achieve better privacy protection effect. Compared with the two existing algorithms, the experimental results show that SARB can achieve better privacy protection effect.

2. Preparation knowledge

2.1. Data structure definition

In the algorithm of this paper, the data structure of quadtree[3] is used to manage the users into different levels and regions. The bottom of the mesh side length is not greater than the threshold l (unit: m). In this experiment, the threshold is set to 35 meters. The threshold is selected according to the user privacy requirements. The larger the threshold is, the higher the user privacy level is. The top of the quadtree is the 0th floor, the bottom is the Hth floor,The i-th grid is 4i. In the entire map, statistics and storage of historical grid points within the distribution of the query points. As shown in Figure 1, the integer in grid (4,2) is 25, which means that the number of historical queries in grid (4,2) is 25 times.

2.2. Query probability

Suppose the map is divided into n $\times$ n location units, the query probability qi of location unit li can be expressed as the ratio of the query times ni of all the users u

3	15	20	16
21	16	25	10
30	7	21	30
30	25	16	5

Fig. 1. Distribution of historical query points

to the location unit to the total query times m of the entire map.

$$qi = \frac{ni}{m} \tag{1}$$

Where i = 1,2, ... ,n^2, and meet$\sum_{i=1}^{n^2} qi$=1.

2.3. Location entropy

Without considering the background information, when the user sends an inquiry request directly to the service provider using the k-anonymous technology, the probability that the service provider deduces the true location of the user is $\frac{1}{k}$,Let pi denote the probability that location loci is the real location of the user.

$$pi = \frac{qi}{\sum_{i=1}^{k} qi} \tag{2}$$

Where i = 1, 2, ..., k, and $\sum_{i=1}^{k} pi = 1$.

In this paper, we use the entropy of location to measure the degree of anonymity, which means that the average uncertainty of the attacker from the anonymous location set to infer the true location of the user. The entropy to obtain the exact position from the candidate position is H.

$$H = -\sum_{i=1}^{K} pilbpi \tag{3}$$

When all k locations in k anonymity technology have the same probability, location entropy H is the largest.

2.4. Normalized distance

We define the normalized distance between the user's real position lr and the i-th dummy position li as di, as shown in equation (4).

$$di = d(lr, li) \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(d-d(lr,li))^2}{2}} \tag{4}$$

d(lr,li) represents the physical distance between the true position lr and the fake position li.

3. System structure

3.1. system structure

he proposed location privacy protection scheme uses a 3-tier entity architecture with a semi-trusted third-party server, As shown in Figure 2. A mobile user is a user set equipped with a mobile intelligent terminal, and generates an anonymous hierarchical grid number of a user according to latitude and longitude coordinates of the user. Semi-trusted third-party server proxy user to generate anonymous area and send the query information, and finally generate accurate query results for the user. The Service Provider (SP) is responsible for finding a candidate result set according to the anonymous query request and returning it to the semi-trusted third-party server.

This article defines the area of the smallest area containing all anonymous locations as an anonymous set cover area.The symbols used in the query process and their definitions are shown in Table 1.

Table 1. Definitions and notations

symbol	definition
h	User-defined anonymous level
ID	User ID
k	User Privacy Protection Level (Greater than or equal to 2)
Amax	The maximum anonymous range parameter set by the user
(R,C)	The coordinates of the user's cell
Con	Query content, this article does not study
(L,U)	User anonymous area upper left cell coordinates
(X,B	User anonymous area lower right cell coordinates

After the user selects the anonymous level h, the cell (R, C) is calculated according to the real position coordinates determined by the positioning function. The user

first sends a service request Qu to a semi-trusted third-party server, The semi-trusted third-party server then constructs the anonymous area and sends the anonymous service request Qs to the service provider.Then the service provider queries according to the request and returns the obtained candidate result set to the semi-trusted third-party server. The semi-trusted third-party server filters the result set according to the grid where the user is located and returns the filtered result to the user. The final user according to their exact location to find the final result. Where the service request Qu = {ID, h, (R, C), k, Amax, Con}, the anonymous service request Qs = {(L, U), (X, B), h, Con}.

3.2. Semi-trusted third-party server architecture

As shown in Figure 2, Semi-trusted third-party server consists of three parts: historical query probability distribution module, anonymous module and query refining module. The history query probability distribution module stores the history query history distribution. After the user submits the anonymous level h, the historical query probability distribution module will use the grid size decided by the anonymous level h as the location unit, and count the number of queries of all location units on the entire map and calculate the query probability, And transmits the historical query probability distribution of the corresponding hierarchical location unit to the anonymous module for the anonymous module to construct an anonymous area for use. Combined with the history query probability distribution and the privacy preference k value and the maximum anonymity interval Amax, the anonymous module generates the corresponding anonymous area using the SARB algorithm (Section 3.3) to obtain the anonymous query Qs, and sends the Qs to the service provider. The query refinement module filters the candidate results returned by the SP according to the location of the user's cell using the anonymous region nearest neighbor search algorithm [4], and sends the filtered result to the user.

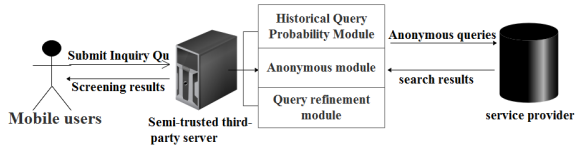

Fig. 2. Architecture of semi-trusted third-party server

3.3. Anonymous location collection selection algorithm

In this paper, the SARB algorithm is used to construct the anonymous region. First, we select 4k location units with similar probabilities to the user's real location according to the background information of the location unit, and then randomly select 2k location units from the 4k grids as the candidate set C, Select k-1 location units from the 2k location units in set C and the real location of the user to form an anonymous area set. To facilitate the description of the algorithm steps, the following definitions of the algorithm's crossover and substitution operations are

given.

Definition 1: Cross operation. Randomly select a cross point in randomly selected position sets Pi and Pj, and cross-connect the data on both sides of the cross point to generate a new set Ps, where i, j, s {1,2,...c_n^{k-1}}. If the selected intersection point is the t-th position unit locit and locjt in Pi and Pj, the two new position sets Pi' and Pj' generated after cross-stitching are {locj1,locj2,... locj(t-1),loci(t+1),...,loci(k-1)} and {loci1

loci2,... loci(t-1),locjt,locj(t+1)... locj(k-1)}.

Definition 2: Replace operation. In a randomly selected location set Pm randomly selected a replacement locmt, and then randomly selected from the set C in a location unit loce, where m∈{1,2,...c_n^{k-1}}, mt∈{m1,m2,...mk-1}, e∈{1,2...n},A new location set Pm ' is generated by replacing the location unit locmt in the set Pm with the location unit loce to{locm1,locm2,... locm(t-1),loce,locm(t+1),... locm(k-1)}.

Anonymous location collection selection algorithm SARB is described as follows:

Input: History query probability distribution, user real location (R, C), user privacy preference k, anonymous level h, the maximum anonymous interval Amax;

Output: User anonymous area Dmax;

step 1: Calculate the query probability q at the location of the user's real location loct;

step 2: Initialize the candidate set C according to the query probability and the background information at the location unit loct where the user's real location is located;

step 3: From the position set C randomly selected Number group position set P,P={P1,P2,...PNumber}, P⊂C, | P | = k-1.

step 4: The last step in the Number group location set {P1,P2,...PNumber} are added to the user's real location loct constitute alternative location set S , S={S1,S2,... SNumber} ,

|S|=Number,Si={loc1,loc2,... lock-1,loct},(i=1,2,... Number), eliminate the set with an anonymous set cover area larger than Amax in location Set S. Filter to get the collection F. The size of the collection F is Psize.

step 5: Calculate the effect of the set Uj of each set of positions in the set F excluding the real position loct on the semantic diversity.

$$scores(Uj) = 1 - \prod_{i=1}^{k-1} \sqrt{2\pi} \frac{di}{d(loct, loci)}$$

j=1,2,... Psize, di denotes the normalized distance between the user's real location loct and the ith false location loci in the anonymous set. d(loct,loci) represents the physical distance between loct and loci. The set of the first Psize.f positions that the set U has a greater influence on the semantic diversity is selected. f is the default survival rate, eliminating the remaining set of positions;

step 6: If the number of iterations reaches the iteration threshold maxiters times, if not reached, the remaining positions after screening in steps 4) and 5) are changed according to the preset crossover probability Pc and the replacement probability Pm to generate a new position Set, and make the number of location set to Number

again; and jump to step 4; if reached maxiters, go to step 7;

step 7: From the set of Psizef locations screened in step 5, a set of locations satisfying argmax (scores (Uj) and the user's real location set loct are selected to form the optimal anonymous set Dmax|D|=k.

Since the query probability q of all location units in Dmax satisfies q1 $\approx$ q2... qk-1 $\approx$ qt. Therefore, all location units in the anonymous location set can ensure the maximum location entropy at the same time. By selecting the optimal set of anonymous locations that meet the requirements of anonymous area, the algorithm maximizes the location entropy and semantic diversity under the premise of ensuring the quality of service, so as to achieve the optimal privacy protection effect.

4. Security Analysis

Because the architecture of this paper is a semi-trusted third-party server architecture, users choose anonymous level h according to their privacy preferences. Therefore, it can be assumed that the semi-trusted third-party server can not obtain the exact location of the user. This section focuses on untrusted service providers, from the probability distribution attacks, homologous attacks and location similarity attacks three major security methods for security analysis.

For the probability distribution attack, the untrusted service provider can extract the anonymous location set Dmax from the query request. Because of the similar query probabilities of location units in the anonymous region constructed in this paper, Therefore, the probability that the location unit loci is the real location of the user is $\frac{1}{k}$.loci$\in$DmaxThat is, the probability that the untrusted service provider deduces the true location of the user from Dmax is still $\frac{1}{k}$. The probability of inferring the true location of the user can not be increased based on the background information.

For homogenous attacks, because the user sets the anonymous level h according to privacy requirements, it does not send its exact location coordinates to any other entity. Untrusted service providers can not get the exact location of a user, even if the user sends multiple location requests consecutively in the same location. Therefore, the program against homologous attacks can effectively reduce the risk of privacy leaks.

For the location similarity attack, the privacy protection scheme in this paper adopts the normalized distance to select the fake location in the candidate location unit to ensure that the distance between any fake location in the location set Dmax and the user's real location loct is not too close, Thus ensuring the formation of an anonymous region contains a variety of semantics. Even if the untrusted service provider acquires all location units of the anonymous location set Dmax according to the query request, it can not deduce the user's privacy information according to the semantic information.

5. Experimental results and analysis

In order to test the privacy protection scheme proposed in this paper, the experiment uses well-known Thomas Brinkhoff road network mobile node data generator to generate simulated mobile object data. Taking the traffic network of 36km×36km in Oldenburg, Germany as an input, 160,773 sampling points generated by 35,763 mobile users are generated by simulation, and the sampling point is used as a query point to initiate a location service query. The data contains user ID, time and geographic coordinates. The threshold of the side length is set to 35 meters, the target area is divided into 1024×1024 grid space, the spatial level is 9 layers, which are 0th to 8th layers respectively.

The experiment compares the proposed SARB algorithm with the dummy algorithm[5]and the Iclique algorithm[6]. dummy is an algorithm for selecting anonymous regions by random walk method without considering query probability. Iclique is a method for implementing position privacy protection based on speed information. In the experiment, two indicators of entropy and semantic types in anonymous area were evaluated.

Figure 3 shows the change of the average entropy value of the anonymous area generated by the three algorithms with the increase of the user privacy preference value k when the number of historical query points is 80,000 and the user anonymous level h = 4. It can be seen that the entropy of anonymous area increases with the increase of user privacy preference k, which indicates that the larger the user privacy preference k is, the larger the entropy value is and the more uncertainty the user's real location is. The SRAB algorithm proposed in this paper is better than the other two algorithms because the anonymous region is constructed in the whole map. It can be seen from the experimental results that the privacy protection scheme in this paper can achieve better privacy protection effect.

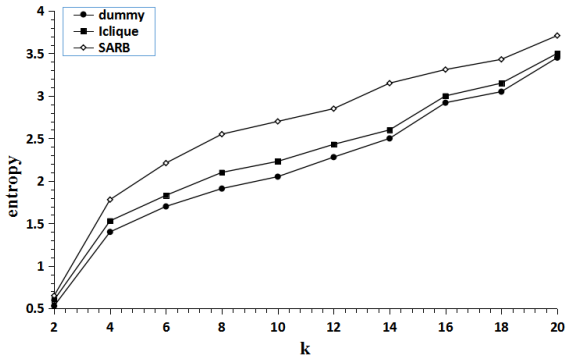

Fig. 3. The entropy value changes with the k value

Figure 4 shows the change of the average entropy value of the anonymous area generated by the three algorithms with the increase of the number of historical query points when the user privacy preference k = 10 and the user anonymous level h = 4. It can be seen that the entropy of the SARB algorithm is obviously larger

than the IClique algorithm and the dummy algorithm. As the number of historical inquiry points increases, the entropy of anonymous regions also increases. The rate of growth slows down and eventually tends to be flat. This shows that the privacy protection scheme designed in this paper, when the number of historical queries reaches a certain amount, affects the privacy protection effect due to the increase of system overhead.

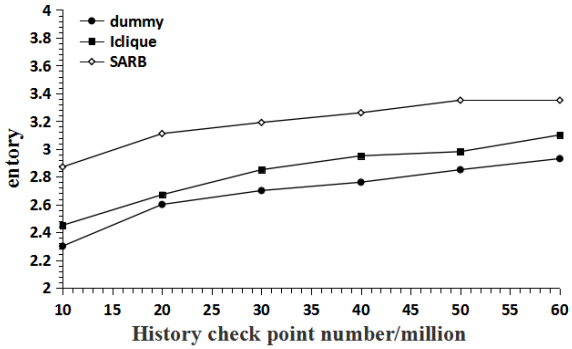

Fig. 4. The entropy value changes with the number of historical query points

Figure 5 shows the change trend of semantic types of anonymous regions generated by the three algorithms as a function of user privacy preference k when the number of historical query points is 60,000 and the user anonymous level h = 3. It can be seen that the anonymous region semantic categories increase with the increase of user privacy preference k. As can be seen from the experimental results, the number of semantic types of SARB is far greater than IClique algorithm and dummy algorithm. Thus can achieve better privacy protection effect. As can be seen from the experimental results, SARB can effectively resist the position similarity attack.

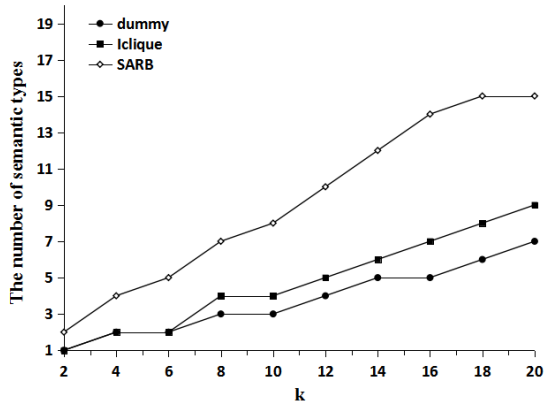

Fig. 5. The number of semantic types varies with k

6. Conclusion

In this paper, an improved method of location privacy protection based on query probabilities is proposed, which uses a semi-trusted server architecture. Users can complete personalized privacy settings according to individual needs and make a choice between the privacy protection effect and service quality. The anonymous region generated by this method has many semantic features and can effectively resist the position similarity attack. The SARB algorithm constructs anonymous regions over the entire map extent, resulting in higher success rates and better anonymity, but at the cost of some overhead. When the anonymity level h and the privacy preference k are both large at the same time, the SARB algorithm may not find enough location units with similar query probabilities. The next step will focus on the SARB algorithm under the condition that the anonymity level h and the privacy preference k are very large application.

References

[1] J. Li, H. Yan, Z. Liu: *Location-sharing systems with enhanced privacy in mobile online social networks.* IEEE Systems Journal *99* (2015), 1–10.

[2] R. Yu, J. Kang, X. Huang: *Mixgroup:accumulative pseudonym exchanging for location privacy enhancement in vehicular social networks.* IEEE Transactions on Dependable&Secure Computing *13* (2016), No. 1, 93–105.

[3] M. F. Mokbel, C. Y. Chow, W. G. Aref: *The new casper:query processing for location services without compromising privacy[C]//International Conference onVery Large Data Bases.* VLDB Endowment (2006), 763–774.

[4] M. Ashouri-Talouki, A. Baraani-Dastjerdi, A. A. Selcuk: *The cloaked-centroid protocol:location privacy protection for a group of users of location-based services.* Knowledge&Information Systems *45* (2015), No. 3, 1–27.

[5] H. Kido, Y. Yanagisawa, T. Satoh: *An anonymous communication technique using dummies for location-based services.* International Conference on Pervasive Services (2005), 88–97.

[6] X. Pan, J. Xu, X. Meng: *Protecting location privacy against location-dependent attack in mobile service.*ACM Conference on Information and KnowledgeManagement (2008) 1475–1476.

Received November 16, 2017