

Research on privacy protection of dummy location interference for Location-Based Service location

International Journal of Distributed
Sensor Networks
2022, Vol. 18(9)
© The Author(s) 2022
DOI: 10.1177/15501329221125111
journals.sagepub.com/home/dsn
 SAGE

Ai Zhang and XiaoHui Li

Abstract

Location privacy refers to the individual private and sensitive location information involved in the user's access to location services. Achieving location privacy protection has become a hot topic of research. However, existing location privacy protection schemes are susceptible to background knowledge attack, edge information attack, and homogeneity attack, on the one hand, and strict constraint on the number of neighbors, on the other hand. To address these deficiencies, a dummy location interference privacy protection algorithm for Location-Based Service location is proposed. To begin with, the dummy location candidate set is constructed based on using WordNet structure to guarantee semantic differentiation, randomly selecting offset location, and conforming to probability similarity; next, the dummy location set is filtered out by discretizing dummy locations based on the Heron formula; finally, the secure anonymity set is constructed according to the anonymity level. Experiments show that the algorithm enhances the privacy protection strength and improves the security of location privacy. Meanwhile, the communication volume and time overhead are reduced and the practicality is boosted by taking into account the sparse and dense environment of location points.

Keywords

Dummy location, LBS, location privacy, privacy protection, security

Date received: 14 March 2022; accepted: 16 August 2022

Handling Editor: Peio Lopez Iturri

Introduction

With the popularity of intelligent devices and the rapid development of mobile terminals, Location-Based Service (LBS) has penetrated and integrated into daily life in all aspects.¹ LBS is an information and utility service that can usually be accessed by mobile devices such as smartphones and global positioning systems that send service requests to LBS servers.² For example, various software such as Twitter, Google Maps, and WeChat are commonly used in life; they bring great convenience and improve the quality of life.³ However, these software services require the user to submit real-time location information that can be analyzed and processed without the user's knowledge, thus

obtaining sensitive information such as the user's home address, workplace, and health condition, which threatens the safety of the user's life and property with

School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou, China

Corresponding authors:

Ai Zhang, School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121000, China.
Email: a15633980578@163.com

XiaoHui Li, School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121000, China.
Email: lhxh@163.com



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work

without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

potential security risk.⁴ Therefore, it is imperative to enhance the security of the user's location privacy.

To address the problem of location privacy protection, a large number of scholars at home and abroad have conducted research and proposed numerous practical methods, which are generally classified into three categories.⁵⁻⁷ Encryption technology⁸ encrypts the user's location information and query request information. Although the quality of service and data availability are guaranteed, the communication volume and time overhead are large; anonymity technology exploits spatial generalization to interfere with the location of the user's query request and uses obfuscated data to obtain the service.⁹ The most typical of this technology is the location K -anonymity technology.¹⁰ However, in the case of sparser users, the requirement to construct an anonymous region containing K users cannot be achieved, and there is a constraint in the number of neighbors;¹¹ distortion technology¹² perturbs the real location information of LBS queries to prevent attackers from directly accessing the real information of users. It includes randomization, dummy location techniques, and so on.

For the dummy location technology, initially proposed by Kido et al.¹³ in 2005, the basic ideology is to add the user's real location to the dummy location and send it to the LBS server together, to confuse the authenticity of the user's location and make it impossible to distinguish the user's real location, thus achieving the user location privacy protection. Lu et al.¹⁴ proposed the GridDummy algorithm based on the virtual squares and the CirDummy algorithm based on the virtual circles. Both algorithms make the dummy locations evenly distributed. It solves the hidden problem that fragmented dummy locations can be easily excluded by attackers and reduces the risk of privacy leakage of user's location due to too small anonymity region. However, both algorithms ignore the possibility that the attacker may refer to background knowledge and exploit edge information, which can easily exclude locations that are partially located in something like the desert, no man's land, and so on, and consequently infer the real location of the user.¹⁵ Arain et al.¹⁶ proposed the MMLPP (multiple mix-zones with location privacy protection) technique to resist background knowledge attack, which is based on both temporal and spatial dimensions, and blocks the attackers from making inference of location privacy information by severing the connection between old and new dummy locations in the way of changing the dummy locations in the mixed region. However, the technique operates much less efficiently in the environment of sparse location points; therefore, it has limitations in terms of the number of neighbors. Niu et al.¹⁷ considered background knowledge and proposed the DLS (dummy-location selection) scheme that uses grids to calculate

the probability of historical location queries to measure the degree of privacy protection through location entropy, but attackers can use techniques such as clustering and data mining to analyze the probability and steal the user's true location information.¹⁸ Therefore, it is not sufficient to consider query probability alone to effectively protect location privacy. Li et al.¹⁹ considered various factors such as historical query probability, speed, and driving direction and proposed a dummy location generation scheme based on user preference selection by selecting dummy locations with high similarity to the user's location to construct an anonymity region. Kamenyi et al.²⁰ proposed a cloud-based architecture to provide authenticated privacy protection with consideration of speed similarity, distance, and other factors. A new privacy-preserving algorithm, AVD-DCA (authenticated velocity-distance based dynamic cloaking algorithm), was designed and implemented, which uses a minimum spanning tree-based dummy location hiding mechanism to protect users' privacy based on their security characteristics and speed similarity. Gustav et al.²¹ proposed a direction-velocity dynamic anonymization algorithm, DSDCA (direction speed dynamic cloaking algorithm), in the case of continuous queries based on location services, which considers similar direction, similar speeds, and the same transmission method for dummy location privacy. The combination of comprehensive factors makes location privacy protection more effective. Although the literature¹⁹⁻²¹ combine more factors to protect users' location privacy to a certain extent, they do not consider the more important location semantic information, which makes it difficult to cope with semantic homogeneity attack²² and easily causes privacy leakage and reduces the utility of the algorithm. Therefore, researchers focus more on improving the quality of service and focus on location semantics. Hara²³ paid attention to location semantics and satisfying user requirements. However, the scheme ignores the dispersion of the generated dummy locations and therefore suffers from the danger of location homogeneity attack.²⁴ Wang et al.²⁵ proposed a maximum and minimum dummy location selection algorithm based on location semantics and query probability, which integrates both location semantics and query probability, and also considers dispersion, quantifying location semantics through computation, but the way of measuring semantic differentiation is crude, the result is not accurate enough, the semantics do not apply widely enough, and there are deficiencies.

In summary, although the above methods protect users' location privacy to a certain extent, they ignore various factors such as location semantics and location dispersion, on the one hand, and are vulnerable to background knowledge attack, edge information attack, homogeneity attack, and so on, which lead to

the leakage of users' location privacy and reduce the security of location privacy protection. On the other hand, the corresponding calculation methods of location semantics and dispersion are crude and have strict constraints for the number of neighbors, which lead to poor execution of the algorithm and reduce the practicality of location privacy protection. Therefore, a dummy location interference privacy protection (DLIP) algorithm for LBS locations is proposed to address the above shortcomings. First, constructing dummy location candidate set under the conditions of semantic diversity, offset location optimization, and probabilistic similarity; second, filtering dummy location set based on the dispersion principle; finally, building secure anonymity set that conforms to the anonymity level.

The DLIP algorithm combines location semantics, offset location, query probability, dispersion, and anonymity level to generate dummy locations that are more reasonable and indistinguishable, which effectively counteracts background knowledge attack and edge information attack; furthermore, the measurement method is relatively accurate, which calculates semantic similarity based on the WordNet structure and discretizes dummy locations through the Heron formula to improve accuracy as well as avoid homogeneity attack; meanwhile, it is universally applicable to environments with sparse and dense location points, which do not have neighborhood constraint and have low operational overhead. The DLIP algorithm is compared with existing algorithms in four aspects: anonymity rate, location entropy, running time, and communication volume, which verifies the effectiveness of the DLIP algorithm for location privacy protection.

Relevant knowledge

Relevant definition

Definition 1 (semantic location). It refers to locations that contain coordinates represented by latitude and longitude, location semantic types, and other features. It has the following characteristics: on the one hand, the sensitivity of the semantic location is defined by the user himself; on the other hand, the same semantics can contain multiple locations and a location can contain multiple semantics.

Definition 2 (WordNet²⁶). It refers to an effective English electronic dictionary containing nouns, verbs, adjectives, adverbs, and so on, with broad lexical coverage and suitable for modern semantic computing. The words in WordNet each form a hierarchical network of synonyms, with each set of synonyms representing a semantic meaning, and it is designed to semantically model words through the categorization of synonyms and existing taxonomic and non-taxonomic

relationships, ultimately constructing a semantic structural form of a tree. It includes a total of 117,659 concept nodes, with a total of 82,115 noun concept nodes, which represents approximately 70% of the total. Therefore, it is very suitable for the measurement of positional semantics involving a large number of noun forms.²⁷

Definition 3 (semantic similarity²⁸). It represents a metric that measures the degree of similarity of the semantics of different locations. Specifically, it means the distance between two words ($T1$, $T2$) at different locations in the WordNet semantic tree. The formula is as follows

$$\begin{aligned} Sim(T1, T2) = & \frac{\sum_{i \in \{1, \dots, |ST1|\}} \max_{j \in \{1, \dots, |ST2|\}} Sim(ST1_i, ST2_j)}{|ST1| + |ST2|} \\ & + \frac{\sum_{i \in \{1, \dots, |ST2|\}} \max_{j \in \{1, \dots, |ST1|\}} Sim(ST2_i, ST1_j)}{|ST1| + |ST2|} \end{aligned} \quad (1)$$

In this formula, Sim denotes the semantic similarity value of two words. $T1$ and $T2$ denote words containing multiple semantics. $ST1_i$, $ST1_j$, $ST2_i$, and $ST2_j$ all denote words containing one semantic meaning. $|ST1|$ and $|ST2|$ denote the number of *senses* of $T1$ and $T2$, respectively, and *sense* refers to the number of semantic categories contained in a single word. According to the formula, the larger the value of semantic similarity, the higher the semantic similarity of the two words, and the smaller the value of semantic similarity, the lower the semantic similarity of the two words.

Definition 4 (Euclidean distance). It represents the distance from location $L_i(X_i, Y_i)$ to another location $L_j(X_j, Y_j)$ and is denoted by $Dis(L_i, L_j)$, in which X and Y denote the latitude and longitude of the location unit, $i \neq j$. The calculation formula is as follows

$$Dis(L_i, L_j) = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \quad (2)$$

Definition 5 (minimum distance). It stands for a criterion that measures the minimum value of the distance between different locations. The calculation formula is as follows

$$Z = \arg \min_{i, j \in \{1, \dots, n\}} \{Dis(L_i, L_j)\} \quad (3)$$

In this equation, Z represents the set of locations that meet the minimum distance principle, L_i and L_j represent any two locations, and Dis represents the Euclidean distance between the two locations.

Definition 6 (query probability). The area around the user is divided into grids, each grid corresponds to a location cell, and the query probability P_i for each location cell is calculated as follows

$$P_i = \frac{N_i}{\sum_{i=1}^n N_i} \quad (4)$$

In this formula, N_i denotes the number of query counts for the location cell corresponding to the grid and $\sum_{i=1}^n N_i$ denotes the sum of query counts for all location cells. In this case, locations with query probability 0 may be unreachable semantic locations such as rivers, deserts, and mountains.

Definition 7 (Heron formula). It is applied to solve the area of a polygon. Based on the theorem that an n -sided shape can be divided into $(n - 2)$ triangles, the calculation of the area of a polygon is transformed into a way of solving the sum of the areas of multiple triangles. The area of a triangle can be found directly using the three side lengths with the following formula

$$S = \sqrt{r(r-a)(r-b)(r-c)} \quad (5)$$

In this formula, S denotes the area of the triangle. r represents the half circumference of the triangle, that is, $r = (1/2)(a + b + c)$, where a , b , and c are the three sides of the triangle.

The dummy locations with a high degree of dispersion are filtered according to the Heron formula. The specific way is shown in Figure 1. Four location points are used as an example, where A denotes the center of the user's real location unit and B, C, and D denote the center of the dummy location units. Ensuring that the locations are relatively discrete is achieved by calculating the maximum value of the quadrilateral area. The Euclidean distance is used to calculate AB, AC, AD, BC, and CD, and thus the area of the two triangles as $S1$ and $S2$, where $S1$ and $S2$ denote the areas of triangles ADC and ABC, respectively. It can be seen that the quadrilateral area is equivalent to the sum of the areas of the two triangles, that is, $S1 + S2$, thus screening out the dummy locations B, C, and D corresponding to the maximum area.

Definition 8 (location entropy²⁹). It is a method of measuring uncertainty, commonly used as a metric to evaluate the degree of location privacy protection and to determine the uncertainty of an attacker identifying the real location.

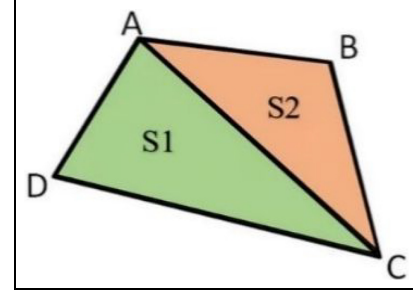


Figure 1. Quadrilateral ABCD.

First, k locations are selected from the grids, and the query probability P_i is calculated for k locations. Then the probability q_i is calculated for each location to be the user's true location, as follows

$$q_i = \frac{P_i}{\sum_{j=1}^k P_j} \quad (6)$$

Finally, the location entropy $H(x)$ is calculated according to q_i , with the following equation

$$H(x) = - \sum_{j=1}^k q_j \times \lg(q_j) \quad (7)$$

It is known that the bigger the location entropy, the lower the probability of the attacker identifying the real location and the stronger the privacy protection effect; conversely, the smaller the location entropy, the higher the probability of the attacker identifying the real location and the weaker the privacy protection effect.

System architecture

The system architecture used for the DLIP algorithm is shown in Figure 2 and consists of three main components: User, Trusted Location Anonymous Server, and LBS Server, of which the Trusted Location Anonymous Server contains the Anonymity Processing Module and the Streamlined Optimization Module. The system architecture has a relatively simple layout, which reduces system overhead and improves efficiency; meanwhile, the application of the Trusted Location Anonymous Server not only improves the privacy strength of location information but also precisely optimizes the results; therefore, it is practical and safe.

The user obtains the map location information of the area through network Wi-Fi, and then sends the location service request and self-defined anonymity level to the Trusted Location Anonymous Server. The

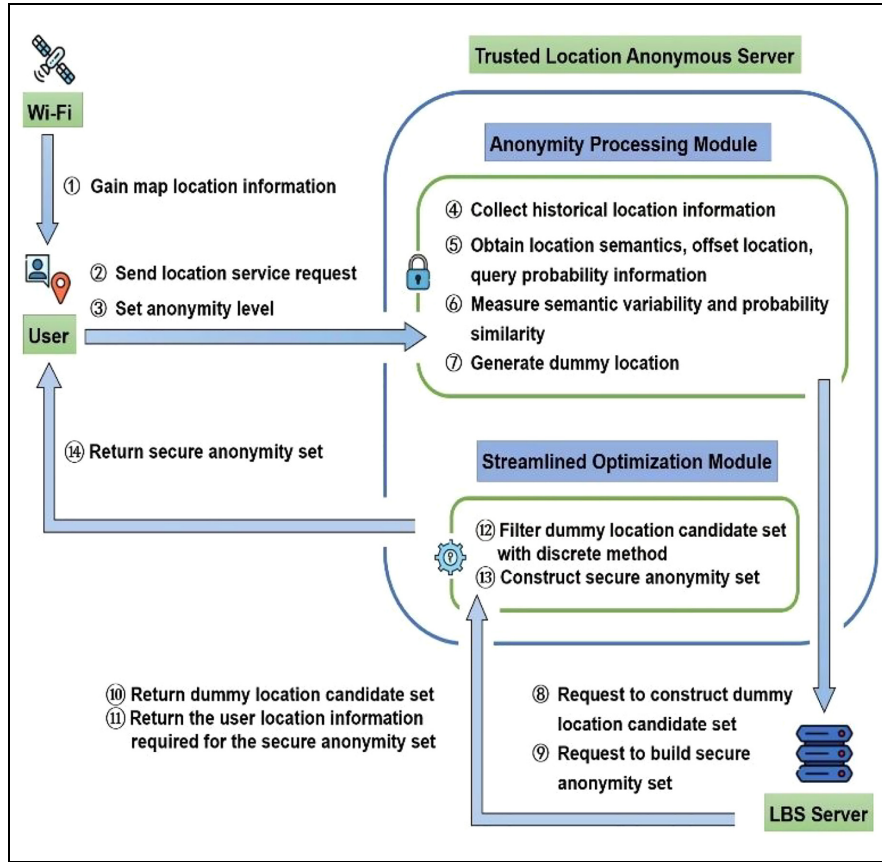


Figure 2. System architecture figure.

Trusted Location Anonymity Server executes the DLIP algorithm to perform dummy location interference to protect location privacy. First, the Anonymity Processing Module obtains historical location, location semantics, offset location, query probability, and other information. It fuzzes the location information, hides key location privacy, and performs dummy location construction; second, it sends a request to the LBS Server to construct a dummy location candidate set and a secure anonymity set; finally, the Streamlined Optimization Module filters the dummy location candidate set returned by the LBS Server based on the discretization principle, and then the secure anonymity set is constructed. In the end, the user's desired secure anonymity set is returned to achieve privacy protection of the user's location.

Privacy protection of dummy location interference algorithms for LBS locations

Algorithm design

The primary ideas of the algorithm are: first, to construct a dummy location candidate set according to the principles; second, to filter the dummy location set

according to the constraints; and finally, to construct a secure anonymity set according to the anonymity level to achieve location privacy desensitization and protect the user's location privacy. The principles for constructing the dummy location candidate set are: measuring the semantic sensitivity of location, implementing offset optimization, and filtering the probability of location query. The constraints for filtering the set of dummy locations are to disperse the dummy locations and to increase the dispersion of the dummy locations. The design of the DLIP algorithm is illustrated in Figure 3.

Based on historical location data, the DLIP algorithm uses dummy location interference technology to achieve location privacy protection. The specific steps are as follows:

Step 1. Obtain historical location information for pre-processing. The historical location points around the user are collected and the location space around the user is divided into grids so that the grids and locations correspond to each other.

Step 2. Construct dummy location candidate set and execute CDLC algorithm. First, the sensitivity of the semantics of the location is measured. The semantic similarity between each location and the user's real

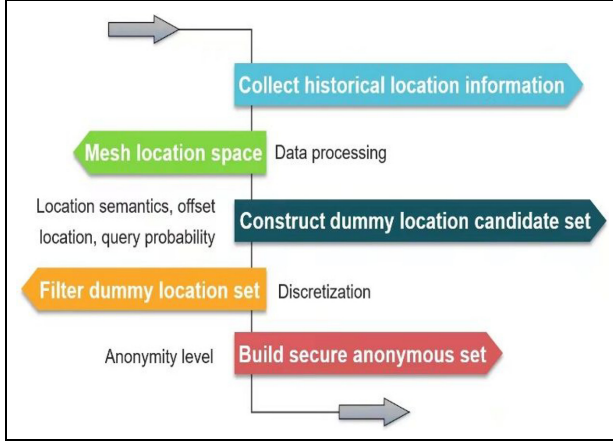


Figure 3. Workflow diagram of the DLIP algorithm.

location is calculated and the result is compared with the sensitivity standard value of $f = 0.5$. If it is greater than or equal to 0.5, it means that the semantics of the location is very similar to the semantics of the user's location, and the location is deleted; otherwise, the location is kept. The final set of remaining locations is denoted as set A . Second, offset optimization is implemented in set A . The set of locations around the user that meet the minimum distance principle is calculated, and the offset location M is randomly selected from the multiple locations that meet the requirements, and the remaining locations are recorded as set B after removing the offset location. Finally, the query probability of the locations in set B is calculated. The query probability of each location can be found by calculating the number of historical queries for each location, and the locations that meet the probability similarity principle are recorded as set C when the query probability of each location is compared with the query probability of the real location of the user.

Step 3. Screen the set of dummy locations and execute the FDLS algorithm. Dummy locations are dispersed by using the Heron formula. The location with the largest Euclidean distance from the user's offset location is selected as the first dummy location and then iterated in order to select the dummy location that corresponds to the maximum area enclosed by the offset location and the identified dummy location; thus, the dummy location with the higher dispersion is filtered out and recorded as set D .

Step 4. Construct secure anonymity set and execute the BSAS algorithm. The final set of dummy locations D is returned and the anonymity level K set by the user is obtained. The set D and the offset

Algorithm 1. CDLC algorithm.

Input: User location TL , historical location dataset H , sensitivity criteria value f
Output: Dummy location candidate set C

```

1.  $n \leftarrow |H|$ 
2. for  $i = 1$  to  $n$  do
3.    $S_i \leftarrow \text{Sim}(TL, L_i)$ 
4.   if  $(S_i \geq f)$ 
5.     delete  $S_i$ 
6.   else insert  $S_i$  into  $A$ 
7.   end if
8. end for
9.  $n \leftarrow |A|$ 
10. for  $i = 1$  to  $n$  do
11.    $V \leftarrow \arg \min_{i \in \{1, \dots, n\}} \{Dis(TL, L_i)\}$ 
12.   Randomly select  $M \in V$ 
13.    $TL \leftarrow M$ 
14. end for
15. Return  $B \leftarrow \text{delete } M \text{ from } A$ 
16.  $n - 1 \leftarrow |B|$ 
17. for  $i = 1$  to  $n - 1$  do
18.    $P(L_i) \leftarrow \frac{N_i}{\sum_{i=1}^{n-1} N_i}$ 
19.   if  $P(L_i) \approx P(TL)$ 
20.     insert  $L_i$  into  $C$ 
21.   end if
22. end for
23. Return  $C$ 

```

location M are combined to form the final secure anonymity set E .

Construct dummy location candidate set

First, basic information such as historical location is collected, and the location information is pre-processed to construct a dummy location candidate set considering three aspects: location semantics, offset location, and query probability, as shown in Algorithm 1.

The inputs to Algorithm 1 are the user location TL , the historical location dataset H , and the sensitivity level criterion f . The ultimate aim is to construct a dummy location candidate set. First, lines 1–8 of the pseudo-code represent the sensitivity of the measured location, thereby ensuring semantic difference of the location. According to equation (1), the semantic similarity S_i between each historical location and the user's real location is calculated. S_i is compared with the standard value f of sensitivity, and if S_i is greater than or equal to the standard value f , it means that the semantics are very similar to the user's sensitive locations, so these locations are deleted; on the contrary, the

Algorithm 2. FDLS algorithm.

Input: Dummy location candidate set C
 Output: Dummy location set D

```

1.  $n \leftarrow |C|$ 
2. for  $i = 1$  to  $n$  do
3.   if  $(i = 1)$ 
4.     Return  $Dis(M, C_i)$ 
5.      $G_1 \leftarrow \max Dis(M, C_i)$ 
6.     insert  $G_1$  into  $D$ 
7.   else
8.     for  $j = 1$  to  $n - 1$  do
9.       Return  $Dis(G_i, C_j)$ 
10.       $G_i \leftarrow S_{max}(\text{Heron formula})$ 
11.      insert  $G_j$  into  $D$ 
12.       $(n - 1) --$ 
13.    end for
14.  end if
15.   $n --$ 
16. end for
17. Return  $D$ 

```

locations that are semantically different from the user's sensitive locations are retained, thus achieving semantic diversity; then, lines 9–15 of the pseudo-code indicate the implementation of offset optimization to select offset location. In set A with different semantics for each location, multiple locations closer to the user's true location are selected according to equation (3) and then together with the user's real location form the offset location candidate set V . Note: The reason for adding the user's real location to the offset location candidate set here is to prevent the situation that there is no suitable candidate location around the user; subsequently, an offset location M is randomly selected in the offset location candidate set to replace the user's real location to construct a secure anonymity set; finally, lines 16–23 of the pseudo-code indicate the calculation of the location query probability to ensure probabilistic similarity. According to equation (4), the location query probability in the set B after removing the offset location is calculated, and the locations with similar probability to the user location query are selected so that the dummy locations are indistinguishable compared to the real location, and the final set of dummy location candidate C is returned.

Filter dummy location set

The set of dummy location candidates C is obtained by Algorithm 1, and the set of dummy locations is discretized by improving the location distribution of set C with the Heron formula, from which the set of dummy locations is filtered, as shown in Algorithm 2.

The input to Algorithm 2 is the set of dummy location candidate C . The purpose is to filter out the dummy locations with high dispersion to form the set

Algorithm 3. BSAS algorithm.

Input: Dummy location set D
 Output: Secure anonymity set E

```

1.  $n \leftarrow |D|$ 
2. for  $i = 1$  to  $n$  do
3.    $K \leftarrow \text{Set SafeRank}$ 
4.   for  $i = 1$  to  $K - 1$  do
5.     insert  $G_i$  into  $E$ 
6.      $E \leftarrow G_i + M$ 
7.   end for
8. end for
9. Return  $E$ 

```

of dummy locations according to the idea of the Heron formula to optimize the location distribution. To begin with, lines 1–6 of the pseudo-code indicate the selection of the first dummy location. The Euclidean distance from each location in the candidate set of dummy locations C to the offset location M is calculated according to equation (2), and the location corresponding to the maximum value is selected as the first dummy location G_1 to be added to the set of dummy locations D ; next, lines 7–17 of the pseudo-code represent the selection of other dummy locations by iteration. The process first calculates the Euclidean distance to G_1 for the remaining locations in the candidate set of dummy locations C excluding G_1 and uses Heron formula (5) to calculate the location corresponding to when M and G_1 form the maximum area to be added as the second dummy location G_2 to the set of dummy locations D . Second, the process calculates the Euclidean distance from the dummy locations in set C to G_2 excluding G_1 and G_2 , and selects the location corresponding to M , G_1 , and G_2 when they form the maximum area as the third dummy location G_3 to be added to the dummy location set D , and so on, until there are no dummy locations in the dummy location candidate set C that meet the requirements. Eventually, the set of dummy locations D is returned.

Build secure anonymity set

Based on the resulting set of dummy location D from Algorithm 2, $K-1$ dummy locations are filtered out based on the anonymity level K and combined with the offset location M to form the secure anonymity set E , as shown in Algorithm 3.

The input to Algorithm 3 is the set of dummy location D . The final output is the secure anonymity set E , which enhances the user's location privacy protection strength. At first, lines 1–5 of the pseudo-code indicate the selection $K-1$ of dummy locations to be added to the secure anonymity set based on the anonymity level set by the user himself; then, lines 6–9 of the pseudo-code indicate the combination of $K-1$ dummy

locations and offset location to return the constructed secure anonymity set E .

Algorithm analysis

Security analysis

1. The authenticity of the user's location is disturbed to effectively counter background knowledge attack. First, the DLIP algorithm integrates location semantics, offset location, query probability, dispersion, and anonymity level to generate more reasonable dummy locations, which effectively reduces the recognition rate of attackers; next, it implements offset optimization to change the idea of applying the user's real location to construct anonymity set in the traditional dummy location algorithm and uses the offset location that is closer to the user to completely replace the user's real location to add to the secure anonymity set, which perfectly conceals the authenticity of the user's location and provides more security; finally, semantic diversity and probabilistic similarity enhance the indistinguishability of the dummy location from the user's real location and improve the strength of privacy protection.
2. The use of historical location information to generate dummy locations successfully prevents the attacker from using geographical knowledge to carry out edge information attack. The attacker can sieve out special locations like mountains, lakes, and deserts when they acquire a certain knowledge background and analyze the set of locations with reduced degree of anonymity, which is likely to infer the real location of the user and cause privacy leakage. The application of historical information in the case of dense or sparse location points can be freed from the number of neighbors, which ensures the number of dummy locations, guarantees the degree of anonymity, and achieves location privacy protection.
3. Semantic differentiation and location discretization are guaranteed to avoid homogeneous attack by the attacker. First, the location that is semantically identical to the user's sensitive location is targeted to deliver useful information to the attacker and expose the user's real location. The removal of locations that are semantically identical to the user's sensitive location is performed first, which fully hides the user's sensitive semantic location, ensures the semantic difference between locations, and avoids semantic homogeneity attack; second, the distribution

of dummy locations in the case of relatively small security anonymity set is too concentrated, and attackers can use the means of location clustering to filter dummy locations and increase the probability of inferring the user's real location, which threatens the user's location privacy. Therefore, it is important to improve the location distribution, select out the scattered dummy locations, and expand the coverage of the secure anonymity set, which can effectively resist the attackers from conducting location homogeneity attack and guarantee the location privacy security of users.

Practicality analysis

1. Measurements of criteria are quantified to improve accuracy. First, the cruder calculation method of traditional location semantics is improved, and the widely applied and accurate WordNet structure is chosen to calculate the semantic similarity. At the same time, the properties of semantics are taken into account, and *sense* is used to mark the case that a location may have multiple semantics, so as to accurately analyze whether the semantics are the same as the user's sensitive location, which makes the measurement results more reliable and more realistic; second, the discrete degree of dummy locations is quantified by applying the Heron formula, Euclidean distance, and so on, which is both convenient and accurate to optimize the location distribution.
2. The communication volume is effectively decreased to improve efficiency. The communication of the algorithm occurs mainly between the User, the Trusted Anonymous Server, and the LBS Server. Since these communications consist of only simple query requests and the transfer of small amounts of data, the communication volume is of constant level and is denoted as $O(C)$. First, the main transmission of location service requests and secure anonymity sets between the User and the Trusted Anonymous Server is $O(C)$, and the transmission of secure anonymity sets depends on the anonymity degree K noted as $O(KC)$; second, the main transmission of dummy location information between the Trusted Anonymous Server and the LBS Server is $O(C)$; meanwhile, it is known to be the first to remove locations with the same sensitive semantics as the User, which can be exempt from operations such as calculating its query probability and whether it is an offset location, thus further reducing the amount of

Table 1. Experimental parameters.

Parameters	Default value	Range of values
Number of trajectories	5719	
Location points	23,718	
K	15	[5, 30]
f	0.5	
Location semantics	Healthcare, public service, leisure and entertainment, education and science, administration and residence, restaurants and shopping centers	
Location sensitive semantics	Hospital, residence	

location information transmission and lowering the communication overhead; finally, in the comprehensive analysis, the overall communication volume complexity of the algorithm will not exceed $O(KC)$ and the communication volume is low, thus having high availability.

3. The time complexity is low to effectively lower the time overhead. The time overhead of the algorithm is mainly concentrated in two stages. In the first stage, the dummy location candidate set is constructed, in which the calculation of semantic similarity and query probability as well as the selection of offset location is reflected in the traversal of locations to achieve the relevant operations, and the time complexity is $O(n)$, so the time complexity of the first stage is $O(n)$. In the second stage, the dummy location set is selected, in which the nested loop of the Heron formula is used to iteratively screen the dummy locations, so the time complexity of the second stage is $O(n^2)$. So overall, the time complexity of the algorithm is $O(n^2)$ and the time overhead is relatively low.

Experiment analysis

Experiment settings

The experiments are performed on the development platform of PyCharm 2019 and the algorithm is implemented through Python programming. The experiments operate on a hardware environment with an Intel Core i7 3.40 GHz processor and a Windows 10 operating system with 16 GB RAM.

The experimental data are generated with the Network-based Generator of Moving Objects proposed and implemented by Thomas Brinkhoff.³⁰ The generator takes as input a traffic map of the German city of Oldenberg (area approx. 16 km \times 16 km). It simulates the movement of the moving objects on the map by setting the number of moving objects, the movement speed, and the running time, and then outputs the location information marked with (object, time, location coordinates), which is used as historical location point

information for the user. In this section, the experiments are conducted on an area of 4 km \times 4 km in the center of the map, which is divided into 1600 grids of 100 m \times 100 m each. The experiments simulate the movement of 200 users in 400-time units at the default movement speed set by the generator, which generates 5719 random trajectories containing 23,718 location points as the dataset for the experiments.

The experimental parameters are set to vary from 5 to 30 for the anonymity level K and 0.5 for the sensitivity criterion f . The experiment classifies location semantics into six types, including healthcare, public services, leisure and entertainment, education and science, administration and residence, and restaurants and shopping centers. The sensitive location semantic type sets are hospital and residence. The specific experimental parameters are shown in Table 1.

The experimental parameters are set, first, for parameter K . The larger the anonymity level K , the corresponding increase in communication volume and running time will reduce the practicality of the algorithm. The smaller the anonymity level K , the smaller the anonymity rate and location entropy, the weaker the privacy protection strength, and the lower the location privacy security. Therefore, regarding the setting of anonymity level K , on the one hand, from the perspective of privacy protection, comprehensive consideration of various factors needs to be carried out and needs to be set as a variable. On the other hand, from the experimental environment, the range of variation can be determined by considering the location points and location semantics in the experimental data set. Therefore, in line with the premise of practical applications, the effectiveness of the DLIP algorithm is verified through the variation of parameter K . The variation range of anonymity level K is set from 5 to 30 to verify the excellent performance of the DLIP algorithm in terms of anonymity rate, location entropy, running time, and communication volume under different anonymity levels, effectively confirming the security and practicality of the DLIP algorithm. In this case, the default value of the anonymity level K is set to 15, on one hand, because the experimental data show that the anonymity rate and location entropy are relatively large

for $K = 15$, which is highly secure in terms of location privacy; meanwhile, the running time and communication volume are relatively low, which is highly practical. The experimental setting of anonymity level $K = 15$ maintains a balance between security and practicality. In the overall view, the experimental comparison is the best at this point, and the DLIP algorithm has the best performance. On the other hand, subsequent comparative experiments have verified that the anonymity level $K = 15$ is optimal, so from a practical point of view, this value, that is, $K = 15$, can be used in practical situations when implementing location privacy protection to optimally solve the location privacy protection problem. Second is for the parameter f . The sensitivity criterion f is a measure of whether other locations are semantically similar to the user's real location. f is set to 0.5 by default because the semantic similarity is a value between 0 and 1, so from a practical application point of view, f is set to 0.5 as a conventional definition criterion to judge the semantic sensitivity of a location. If the semantic similarity is greater than or equal to 0.5, it means that the location is semantically similar to the user's real location, which is sensitive and needs to be deleted in advance. Conversely, if the semantic similarity is less than 0.5, it means that it is different from the user's real location semantics and needs to be retained. Subsequent experiments based on the above situation can be compared to verify the effectiveness of the DLIP algorithm, which in turn can verify the reasonableness of the parameter K and f settings.

Experiment comparison analysis

The experiments are mainly to verify the effectiveness of the DLIP algorithm for location privacy protection. The comparison experiments measure the strength of location privacy protection by four metrics: anonymity rate, location entropy, running time, and communication volume. The ARB (anonymous region building) algorithm³¹ and the MMDS (maximum and minimum dummy selection) algorithm are chosen to compare with the DLIP algorithm to verify the best utility of the DLIP algorithm.

For the comparison algorithms, first, the ARB algorithm is a location privacy protection method based on query probability, which combines query probability information to generate user anonymous regions, thus resisting edge information attack by attackers and ensuring user's location privacy security. The ARB algorithm is chosen for comparison mainly because it only considers a single factor of query probability and ignores various important factors such as location semantics, offset location, query probability, dispersion, and anonymity level, which makes it only resistant to edge information attack and unable to cope with background knowledge attack and homogeneity attack; thus, it is easy to cause user's location privacy

leakage. In this way, the DLIP algorithm highlights the importance of considering multiple factors and the security against background knowledge attack, edge information attack, and homogeneity attack. It also improves the practicality and highlights the excellent performance of the DLIP algorithm in terms of security. Second, the MMDS algorithm is a dummy location selection algorithm based on location semantics and query probability. Its combination of semantic information, query probability, and location dispersion factors avoids attackers combining background knowledge to filter dummy locations, which improves location privacy security. The MMDS algorithm is chosen for comparison mainly because although it considers multiple factors and improves security to a certain extent, the measurement of location semantics and location dispersion is crude and has limited applicability, as well as the algorithm has high volume of communication and time overhead, which has drawbacks in terms of practical application. In this way, the accuracy of the DLIP algorithm in quantitatively calculating location semantic similarity and measuring location dispersion using the Heron formula is highlighted. The algorithm not only ensures security but also reduces the communication volume and time overhead, and highlights the excellent performance of the DLIP algorithm in terms of practicality.

Anonymity rate. The anonymity rate³² is used to reflect the degree of location privacy protection of the algorithm in the presence of attacks. The same experimental environment is first created so that the algorithms can run on the same dataset and consistent parameter settings for reasonable comparisons. A typical Bayesian mechanism with attacks such as background knowledge and homogeneity is applied here.³³ The experimental result is shown in Figure 4.

The experimental results show that, first, the anonymity rate of all three algorithms decreases against the same attack. This is because as the anonymity level increases, the number of dummy locations satisfying the requirement decreases accordingly, which makes the construction of the secure anonymity set more difficult and leads to the decrease of the privacy protection effect. Second, the DLIP algorithm that combines various factors such as location semantics, offset location, and query probability has the highest anonymity rate and is the highest when the anonymity level is $K = 15$. The DLIP algorithm has an overall high and stable anonymity rate; thus, it can effectively resist attacks such as background knowledge and homogeneity, which provides excellent privacy protection. The MMDS algorithm is second only to the DLIP algorithm in terms of anonymity rate, which gradually increases when $K < 15$ and decreases when $K \geq 15$.

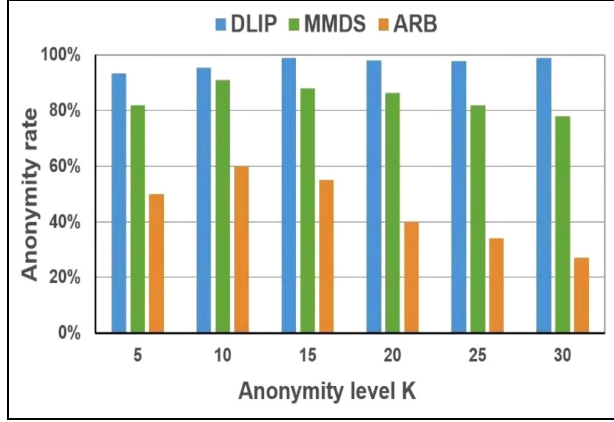


Figure 4. Comparison figure of anonymity rate under Bayesian attack.

This is because the MMDS algorithm is not accurate enough in measuring semantics and dispersion, the number of locations at the beginning is small, and the disadvantages are not obvious, so the anonymity rate is better. However, as the anonymity level increases, it becomes more difficult to generate dummy locations and the anonymity rate gradually decreases. The ARB algorithm has the lowest overall anonymity rate, which gradually increases when $K < 15$ and gradually decreases with an obvious downward trend when $K \geq 15$. It indicates that if query probability alone is considered, and the location semantics and dispersion, and so on are ignored, it will make the correlation between dummy locations enhanced and more vulnerable to background knowledge attack, and so on, which makes the location privacy less secure. In summary, it can be verified that when $K = 15$, the DLIP algorithm has the highest anonymity rate, the strongest security, the best performance, and the best comparative experimental results at this time, so the anonymity level $K = 15$ can be applied to the field of location privacy protection in practical situations, which has important practical significance.

Ideally, the location points appear sparse, but in reality, they are mostly dense. To highlight the wide range of applicability of the DLIP algorithm and get rid of the constraint of the number of neighbors, the experiment compares the anonymity rate for the number of different locations in the region. The experimental result is shown in Figure 5.

The experimental comparison shows that the anonymity rate of all three algorithms decreases as the number of locations increases. First, the DLIP algorithm always has the highest anonymity rate and is the most adaptable, regardless of whether the location points are sparse or dense. Second, the MMDS algorithm has the lower anonymity rate than the DLIP algorithm and decreases more than the DLIP algorithm because the

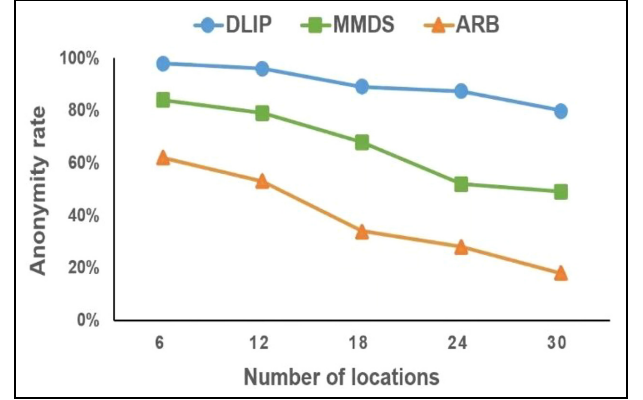


Figure 5. Comparison of anonymity rate for the number of different locations.

semantic measure is coarser, which affects the utility of the dummy locations and decreases the security of location privacy protection. Finally, the ARB algorithm has the lowest and substantially lower anonymity rate, and the worst location privacy protection due to semantic similarity and density between dummy locations, which raises the risk of homogeneity attack.

Location entropy. The strength of privacy protection is measured by applying the location entropy from the literature,²⁹ which measures the effectiveness of location privacy protection. The experiment comparison with different anonymity levels is shown in Figure 6.

The experimental results show that, first, as the anonymity level K increases, the location entropy of the three algorithms shows an overall trend of increasing, and the strength of location privacy protection increases. Moreover, when $K < 15$, the growth trend of all three algorithms is more obvious, and when $K \geq 15$, the growth trend of all three algorithms is flatter, so the comparison experiment is remarkable when $K = 15$. Second, the location entropy of the DLIP algorithm is significantly greater than the other two algorithms because the DLIP algorithm takes into account the location semantics, dispersion, offset location, and other factors and measures them accurately, which enhances the uncertainty of identifying the user's real location and can protect the user's location privacy more effectively, thus achieving the best location privacy desensitization.

Running time. The time overhead of the DLIP algorithm is mainly reflected in the removal of locations with the same semantic sensitivity as the user, the selection of offset location, the calculation of semantic similarity, the calculation of query probability, and the measurement of dispersion. The variation in running time of

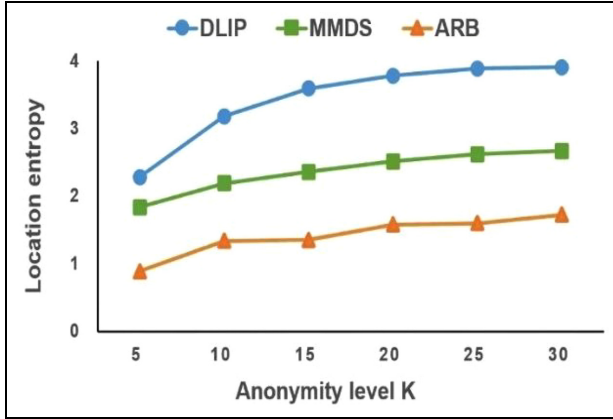


Figure 6. Location entropy comparison figure.

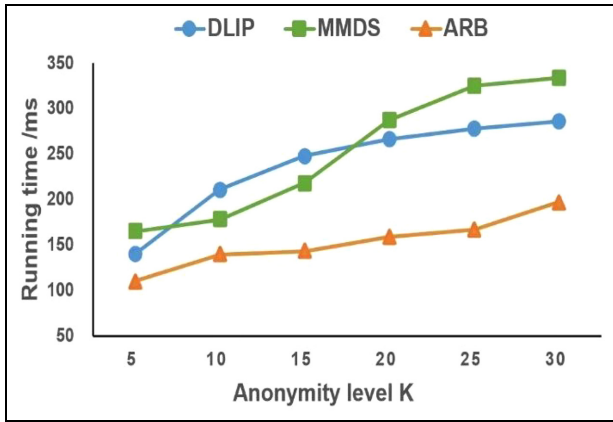


Figure 7. Running time comparison figure.

the three algorithms is compared for different anonymity levels, as shown in Figure 7.

The experimental results show that the running time of all three algorithms becomes longer as the anonymity level K increases. The ARB algorithm does not take into account location semantics and dispersion, so it has fewer operations and minimal running time. In contrast to the MMDS algorithm, the DLIP algorithm initially performs the deletion of sensitive semantic locations, and the number of historical locations is reduced accordingly, so the running time is smaller. However, the time overhead of using the Heron formula for dummy location screening is larger, and the running time is higher than the MMDS algorithm to some extent. As the anonymity level K increases, the time overhead of the DLIP algorithm grows relatively slowly; however, the time overhead of the MMDS algorithm grows more obviously, which eventually makes the time overhead of the DLIP algorithm gradually smaller than that of the MMDS algorithm. In particular, when $K = 15$, the growth trend of the MMDS algorithm and DLIP algorithm starts to change

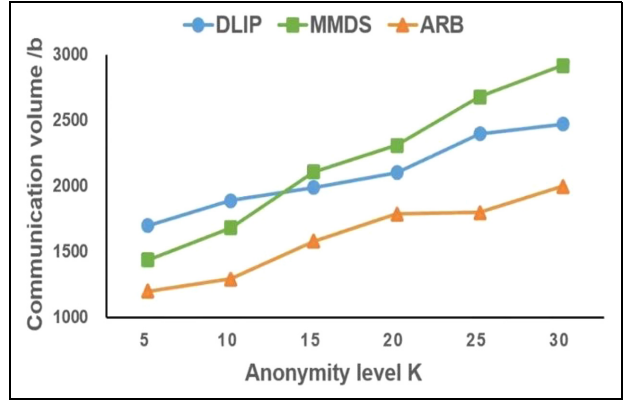


Figure 8. Communication volume comparison figure.

significantly, and the experimental comparison effect is outstanding. Overall, the DLIP algorithm is more advantageous in terms of running time, more efficient, and more practical.

Communication volume. The amount of communication is mainly reflected in the construction of the dummy location candidate set and the screening of the dummy location set. The amount of work reflects the amount of communication and the practical performance of the algorithm. The change in communication volume is observed by setting the value of the anonymity level K , as shown in Figure 8.

The experimental result shows that as the anonymity level K gradually increases, the communication volume of all three algorithms shows an increasing trend. First, the ARB algorithm has the single consideration and the smallest communication volume. Second, it can be seen that when $K \leq 13$, the communication volume of the MMDS algorithm is smaller than that of the DLIP algorithm, because the DLIP algorithm performs operations such as deleting sensitive semantic locations and selecting offset locations. However, when $K \geq 13$, the MMDS algorithm increases significantly more than the DLIP algorithm. It can also be seen that when $K = 15$, the DLIP algorithm communication volume is not only lower than the MMDS algorithm but also in the state of less communication volume when the DLIP algorithm practical performance is optimal. Therefore, the DLIP algorithm has an overall smaller communication volume and more obvious advantage to ensure the quality of service.

By concluding the above comparative experiments, the DLIP algorithm not only implements offset optimization, but also takes into account semantic differentiation and dispersion of dummy locations. It has a high anonymity rate and location entropy, which can improve the strength of privacy protection and enhance the security of location privacy protection. Meanwhile,

it decreases communication and time overhead, which improves efficiency and practicality. Therefore, the DLIP algorithm can effectively achieve location privacy protection with high utility.

Conclusion

A dummy location interference privacy protection algorithm DLIP for LBS location is proposed to address the deficiencies of traditional techniques with the strict constraint on the number of neighbors, as well as to effectively avoid background knowledge attack, edge information attack, and homogeneity attack.

The algorithm integrates five aspects of location semantic sensitivity, offset location, query probability similarity, location distribution discretization, and compliance with anonymity level to achieve location privacy desensitization.

The core idea of the algorithm is divided into three parts: the construction of the dummy location candidate set, the screening of the dummy location set, and the construction of the secure anonymous set. First, the construction of the candidate set of dummy locations is achieved according to three principles: semantic differentiation of locations, the optimal substitution of offset location, and similarity of query probability; second, according to the criterion of high dispersion of dummy locations, the Heron formula is used to screen dummy locations and eventually construct more reasonable and indistinguishable ones; finally, the offset location is combined into secure anonymity set to enhance the strength of location privacy protection and achieve the effectiveness of location privacy protection.

The experimental analyses compare the DLIP algorithm with the MMDS algorithm and ARB algorithm based on four metrics: anonymity rate, location entropy, running time, and communication volume. The experimental results show that the DLIP algorithm balances security and practicality; it not only achieves the effectiveness of location privacy protection but also guarantees the quality of service, which has excellent practical application value.

Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Natural Science Foundation of China Youth Fund Grant

(61802161), the Liaoning Provincial Education Department Scientific Research Funding Project (JZL202015402).

ORCID iD

XiaoHui Li  <https://orcid.org/0000-0003-2357-1642>

References

1. Peng T, Liu Q, Wang G, et al. Multidimensional privacy preservation in location-based services. *Future Gener Comp Sy* 2019; 93: 312–326.n
2. Shaham S, Ding M, Liu B, et al. Privacy preservation in location-based services: a novel metric and attack model. *IEEE T Mobile Comput* 2020; 20(10): 3006–3019.
3. Sun G, Cai S, Yu H, et al. Location privacy preservation for mobile users in location-based services. *IEEE Access* 2019; 7: 87425–87438.
4. He X, Jin R and Dai H. Leveraging spatial diversity for privacy-aware location-based services in mobile networks. *IEEE T Inf Foren Sec* 2018; 13(6): 1524–1534.
5. Jiang H, Li J, Zhao P, et al. Location privacy-preserving mechanisms in location-based services: a comprehensive survey. *ACM Comput Surv* 2021; 54(1): 1–36.
6. Zhang QY, Zhang X, Li WJ, et al. Overview of location trajectory privacy protection technology based on LBS system. *Appl Res Comput* 2020; 37(12): 3534–3544.
7. Ye H, Han K, Xu C, et al. Toward location privacy protection in spatial crowdsourcing. *Int J Distrib Sens N* 2019; 15(3): 1550147719830568.
8. Chen G, Zhao J, Jin Y, et al. Certificateless deniable authenticated encryption for location-based privacy protection. *IEEE Access* 2019; 7: 101704–101717.
9. Ma C, Zhou C and Yang S. A Voronoi-based location privacy-preserving method for continuous query in LBS. *Int J Distrib Sens N* 2015; 11(3): 326953.
10. Fei F, Li S, Dai H, et al. A K-anonymity based schema for location privacy preservation. *IEEE Trans Sustain Comput* 2019; 4(2): 156–167.
11. Zhang S, Li X, Tan Z, et al. A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services. *Future Gener Comp Sy* 2019; 94: 40–50.
12. Wu Z, Wang R, Li Q, et al. A location privacy-preserving system based on query range cover-up or location-based services. *IEEE T Veh Technol* 2020; 69(5): 5244–5254.
13. Kido H, Yanagisawa Y and Satoh T. Protection of location privacy using dummies for location-based services. In: *21st international conference on data engineering workshops (ICDEW'05)*, Tokyo, Japan, 3–4 April 2005, pp.1248–1248. New York: IEEE.
14. Lu H, Jensen CS and Yiu ML. PAD: privacy-area aware, dummy-based location privacy in mobile services. In: *Proceedings of the seventh ACM international workshop on data engineering for wireless and mobile access*, Vancouver, BC, Canada, 13 June 2008, pp.16–23. New York: ACM.
15. Argyros G, Petsios T, Sivakorn S, et al. Evaluating the privacy guarantees of location proximity services. *ACM Trans Priv Secur* 2017; 19(4): 1–31.

16. Arain QA, Memon I, Deng Z, et al. Location monitoring approach: multiple mix-zones with location privacy protection based on traffic flow over road networks. *Multimed Tools Appl* 2018; 77(5): 5563–5607.
17. Niu B, Li Q, Zhu X, et al. Achieving k-anonymity in privacy-aware location-based services. In: *IEEE INFOCOM 2014-IEEE conference on computer communications*, Toronto, ON, Canada, 27 April–2 May 2014, pp.754–762. New York: IEEE.
18. He W. Research on LBS privacy protection technology in mobile social networks. In: *2017 IEEE 2nd advanced information technology, electronic and automation control conference*, Chongqing, China, 25–26 March 2017, pp.73–76. New York: IEEE.
19. Li C, Zhang X, Yan F, et al. False location generation scheme based on user preference selection. *Comput Eng Des* 2019; 40(4): 914–919.
20. Kamenyi DM, Wang Y, Zhang F, et al. Authenticated privacy preserving for continuous query in location based services. *J Comput Inf Syst* 2013; 9(24): 9857–9864.
21. Gustav YH, Wang Y, Domenic MK, et al. Velocity similarity anonymization for continuous query location based services. In: *2013 International conference on computational problem-solving (ICCP)*, Jiuzhai, China, 26–28 October 2013, pp.433–436. New York: IEEE.
22. Chaaya KB, Barhamgi M, Chbeir R, et al. Context-aware system for dynamic privacy risk inference: application to smart IoT environments. *Future Gener Comp Sy* 2019; 101: 1096–1111.
23. Hara T. Dummy-based location anonymization for controlling observable user preferences. In: *2019 IEEE global communications conference (GLOBECOM)*, Waikoloa, HI, 9–13 December 2019, pp.1–7. New York: IEEE.
24. Jiang H, Zhao P and Wang C. RobLoP: towards robust privacy preserving against location dependent attacks in continuous LBS queries. *IEEE/ACM T Network* 2018; 26(2): 1018–1032.
25. Wang J, Wang CR, Ma JF, et al. Dummy location selection algorithm based on location semantics and query probability. *J Commun* 2020; 41(3): 53–61.
26. AlMousa M, Benlamri R and Khoury R. Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet. *Knowl-Based Syst* 2021; 212: 106565.
27. Zhu X, Yang X, Huang Y, et al. Measuring similarity and relatedness using multiple semantic relations in WordNet. *Knowl Inf Syst* 2020; 62(4): 1539–1569.
28. Li F, Liao L, Zhang L, et al. An efficient approach for measuring semantic similarity combining WordNet and Wikipedia. *IEEE Access* 2020; 8: 184318–184338.
29. Sun Y, Chen M, Hu L, et al. ASA: against statistical attacks for privacy-aware users in location based service. *Future Gener Comp Sy* 2017; 70: 48–58.
30. Brinkhoff T. A framework for generating network-based moving objects. *Geoinformatica* 2002; 6(2): 153–180.
31. Zhao DP, Song GX, Jin YY, et al. Query probability-based location privacy protection approach. *J Comput Appl* 2017; 37(2): 347–351.
32. Zhao P, Liu W, Zhang G, et al. Preserving privacy in WiFi localization with plausible dummy locations. *IEEE T Veh Technol* 2020; 69(10): 11909–11925.
33. Shokri R, Theodorakopoulos G, Troncoso C, et al. Protecting location privacy: optimal strategy against localization attacks. In: *Proceedings of the 2012 ACM conference on computer and communications security*, Raleigh, NC, 16–18 October 2012, pp.617–627. New York: ACM.