Location Privacy Preservation Mechanism for Location-Based Service With Incomplete Location Data

XUDONG YANG1, LING GAO1,2, JIE ZHENG1, AND WEI WEI3, (Senior Member, IEEE)
1School of Information Science and Technology, Northwest University, Xi’an 710127, China
2School of Information Science and Technology, Xi’an Polytechnic University, Xi’an 710048, China
3School of Computer Science, Xi’an University of Technology, Xi’an 710048, China
Corresponding authors: Ling Gao (gl@nwu.edu.cn) and Wei Wei (weiwei@xaut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672426, Grant 61572401, and Grant 61701400, in part by the Key Research and Development Program of Shaanxi Province under Grant 2018ZDXM-GY-036, and in part by the National Key Research and Development Project under Grant 2019YFC1521400.

ABSTRACT The Location-Based Service has been widely used for mobile communication networks and location systems. However, privacy disclosure for incomplete collection location data in LBS was ignored in most of the existing works. To solve the problem of privacy disclosure, we propose a location privacy method based k-anonymity to prevent privacy disclosure in LBS constrained in incomplete data collection. The proposed scheme can provide effectively location privacy-preserving in the process of constructing the anonymous set, and against background attacks. In this method, we first designed a construct method for anonymous candidate set (ACS) with compressing sensing technology, to solve the problem of incomplete data of collection location. To prevent the privacy disclosure in the process of construct anonymous, we then adapt the differential privacy mechanism to construct the anonymous set (AS) with the ACS. We finally used the optimization method based on the Stackelberg game model to improve the privacy level of AS to against probabilistic attack. As shown in the theoretical analysis and the experimental results, the proposed method can achieve significant improvements in terms of privacy degree and applicability.

INDEX TERMS Location privacy, LBS privacy, K-anonymity.

I. INTRODUCTION

In recent years, with the rapid development of mobile internet technology and smart mobile devices, people can obtain various Location-Based Services (LBS) through mobile smart devices. The LBS can meet people’s service requests and improve service efficiency by providing the nearby service based on the user’s request location. Many high LBS practical applications can facilitate people’s life such as Dianping, Gaode Map, Meituan Takeout, etc. These applications obtain the location of the user’s service request at the central server and then the server returns corresponding results of user’s requested.

Although LBS provides several benefits to people’s lives, it still raises serious privacy issues. For example, when people send their service requests to the central server of LBS, the attacker can obtain the user’s request location due to the un-trust server of LBS or the attack aim at the server of LBS. In this way, the sensitive locations, which store in the server of LBS, can be stolen by the attacker. The sensitive location of users refers to the location like the hospital, home, and company that can reflect the user’s behavior habits and mobility pattern who were requested. When the disclosure of these locations happening, they can help the attackers to get sensitive information of users and thus infer user’s roles, behaviors, and habits [1]. This would lead to the disclosure of the user’s privacy and hindered the popularity of LBS.

To address the privacy disclosure issues mentioned above, researchers have proposed many methods to preserve location privacy in LBS [31]–[33]. The key idea of these methods is to avoid direct requests to the server of LBS from real request locations. Broadly speaking, existing works of location privacy-preserving can be classified into three categories: offset method [5]–[7], dummy method [2], [3], [8]–[11] and spatial cloaking method [4], [12]–[15]. The offset method is easy to implement but has poor availability. In the dummy and spatial cloaking methods, the k-anonymity idea has a wide range of applications due to its good balance between privacy.
and availability. The dummy method realizes privacy protection through anonymity region which contains \(k-1\) dummy locations and one request location. This approach has a high degree of privacy because of the anonymous locations from the dummy location. The spatial cloaking achieves privacy protection of location and high availability by constructing the cloaked region includes at least \(k-1\) other user’s locations around the location of the requested user.

Although the above methods based on k-anonymity can prevent privacy disclosure in LBS, there remain three challenges. Firstly, the incomplete collection locations. It causes incomplete collection locations due to anonymous server failure or other reasons. It is easy to lead to the anonymous failure problem that candidate location was less than \(k-1\) because of the missing collection locations. Secondly, the attacker’s background knowledge. The request location of user in the anonymous set can be easily obtained by ignoring the attacker’s background knowledge. The attacker can combine the user’s request location probability distribution and the background knowledge such as the protection algorithm to infer the user’s real location in the anonymous set. Thirdly, the privacy disclosure when we constructing the anonymous set. The attacker can determine the anonymous locations, which were added in the anonymous set by monitoring the change of anonymous candidate set in the anonymous server and, reduce the difficulty of his attack. All of these challenges will result in the loss of user’s privacy.

In this paper, we proposed a novel k-anonymity location privacy through the generation of the anonymous candidate set, construction of anonymous set and optimization of anonymous set scheme to address the aforementioned challenges. Specifically, our scheme adopts the k-anonymity technique, combined with the compressed sensing technology, differential privacy and game model to preserve users’ location privacy. In our method, we first use the compressed sensing technology to completing the missed location of collected and then construct the anonymous candidate set(ACS) with the collected data after completing. Then, to avoid privacy disclosure in this process of constructing the anonymous set, we use the exponential mechanism of differential privacy to choosing the anonymous locations from the ACS. Finally, we adopt the Stackelberg model between attack strategy and protection strategy to optimizing the anonymous set and against the background attacks. Overall, our contributions in this paper are as follow:

1. We design a k-anonymity location privacy protection method for incomplete location data in LBS, which not only effectively prevents user request location privacy disclosure after its complete construction, but also prevents location privacy disclosure during anonymous collection construct.

2. We propose a novel anonymous construction method to prevent privacy disclosure in this process with Differential Privacy. Moreover, this method provides strong protection against probabilistic attacks.

3. Considering the background attack, we present an optimization method during anonymous set construction based on the Stackelberg Game model between users and attackers, to defensive background attack. Finally, the proposed method is evaluated through extensive experiments on a real-world dataset.

The rest of this paper is organized as follows. Section II reviews the related work. Section III gives Motivation and Basic idea. Following in Section IV, we describe the details of our method. Then, we analyze the security of our method in Section V. Finally, we evaluate the performance and draw the conclusion in Section VI and VII, respectively.

II. RELATED WORK

In recent years, a lot of works have been done to protect the privacy of user’s request location in LBS, including the offset dummy and spatial concealment. Also, spatial concealment and dummy technology borrow the idea of k-anonymity to protect location privacy.

The offset reduces attacker’s infer accuracy by adding the noise of location to achieve the effect of protecting user privacy, such as using the small area represent the real location. The SpaceTwist achieved location offset mainly by setting anchors around the request location and send services request of LBS based on the anchor location [5]. In reference [6], an offset vector distribution method was proposed to constrain the offset location distribution and to resist the statistical probability attack when the probability density of migration location is constant. By using the differential privacy mechanism, Cao et al. proposed a position offset method to resist the time correlation of user’s position, and provided the disturbed location with differential privacy to improve the location availability while resisting attacks [7].

As an important method to protect location privacy, the anonymous method with dummy locations, called dummy, has attracted researchers’ attention. The dummy method generates the anonymous set consists of \(k-1\) dummy locations and one real requested locations to achieve privacy protection. The PAD, use the virtual grid to generate the desired dummy locations and to construct an anonymous collections using these dummy locations [8]. In order to further improve the efficiency of defense, Zha et al. proposed a method using virtual circles and grids to generate dummy location to constructed anonymous collections [5]. Niu et al. used the location entropy for generating dummy locations that meets the query probability constraint and realized the construction of anonymous collection [9]. At the same time, in order to save the cache space and improve the hit ratio, Niu et al. generated dummy locations through the method of anonymous entropy and improved the protection effect of anonymity by combining the hit ratio and maximizing the privacy of the query [10].

Although the dummy can address privacy issues successfully, its low availability and high computation expense hinders practices application. Thus, the anonymous method which called spatial concealment, which anonymous locations from other user’s request locations. The spatial concealment technology constructs an anonymous area containing \(k\)
real location for each query user and making it difficult for service providers to determine the real request and against background attack. Literature [11] proposed a privacy method to construct anonymous areas by using the Hilbert curve. In detail, the users location were sorted by the Hilbert curve, and k adjacent location were divided into groups, the anonymous areas were constructed by the same group of location. Literature [12] combined the method of constructing anonymous region with differential privacy and Hilbert curve, and perturbed the similar region with the same Hilbert curve. In this way, the smaller the anonymous region is, the smaller the position can be regarded as the disturbance result. Many other researchers used the entropy of position to divide their positions with similar historical query probability into the same anonymous region to realize k-anonymity [13], [14].

Different from existing works, the proposed scheme achieves k-anonymity by carefully generating real locations surrounding request location and under the missed location. It can not only protect privacy after the completion of anonymous set but also provide privacy security in the process of anonymous set construction.

III. MOTIVATION
The anonymous set consists of a user’s request location and the number of k − 1 anonymous locations. The anonymous locations in this paper is mainly derived from the other user’s locations around this user’s request location. Generally, the process of construct anonymous set works through the following two steps:

Step1: The anonymous candidate set is constructed by selecting the real location near the request location;
Step2: The anonymity set is constructed by selecting the anonymous location from anonymous candidate set.

In the construction of the anonymous candidate set, existing works construct the anonymous candidate set with the locations, which from the collection of the anonymous server. If the anonymous server can run safely and smoothly, this method can construct successfully. Otherwise, because of the lack of effective location data, which can lead to the failure of construct anonymous candidate sets. This is due to incomplete collection data in the anonymous server and the incomplete collection data stem from the following two aspects: On one hand, the server failure causes the location loss in the process of location acquisition. In this case, when the number of candidate anonymous locations around the real environment is less than k − 1, it is difficult to meet the requirement of k-anonymity. On another hand, the continuous stay of the surrounding users in one location also leads to fewer available anonymous candidate locations. In this situation, the user has to choose a further location for anonymity, which increases the anonymous cost. In summary, incomplete data collection will increase the cost of anonymous and even lead to failure anonymous.

In the construction of the anonymous set, most researchers ignore the issue of privacy disclosure involved in choosing anonymous locations. The adversary can infer the anonymous location by stealing the selection probability of the location in the candidate collection. In particular, the occurrence of the following two situations to lead to anonymous construct with privacy disclosure: 1) The attacker can reduce the difficulty of the attack by filtering out anonymous locations with a low probability of historical requests; 2) The attacker can infer the real request location from the difference in distribution probability of the historical anonymous collection. Because the location data selection in the process of anonymous collection construction causes the occurrence of the above problems, we regard this situation as a privacy leakage problem in the process of anonymity. That above attack named the probability attack to refer to that the adversary can obtain the real requested location not only through the different anonymous location request probability in the anonymous set at a single time but also through the difference of the anonymous set distribution in different request time.

After the construction of the anonymous set, the adversary can still identify real request locations according to his background knowledge in the prior works. The background knowledge of the adversary is the historical protection mechanism. From this knowledge, the adversary can infer the location that the user may appear based on the Bayes theorem and filter out some anonymous locations. Thus the attacker can reduce the privacy level of the constructed anonymous set.

To address the aforementioned problems, the main purpose of our scheme is to design an efficient scheme, which fully considers missing location in the constructing anonymous candidate set, privacy disclosure in constructing the anonymous set and the background attack mentioned above. Thus, there are three steps in our method:

(1) Construction of anonymous candidate set: To solve the problem of incomplete collection data, we propose an construction scheme of anonymous candidate set by integrating the Compressed Sensing technique and the Realistic Constrain. we can obtain the complete collection dataset with reconstructing the missed request data of user location, which by the Compressed sensing technology. Then, the candidate data set is constructed by choosing the locations that satisfying the realistic constraints from the complete collection data.

(2) Construction of anonymous set: To prevent the leakage privacy of locations selection, we propose an construction of anonymous set with differential privacy. It can cloak location selection probability by using the exponential mechanism of differential privacy. At the same time, the ability to against background attack is improved, by designing the scoring function of the exponential mechanism that takes into account to resist the single-moment attack and continue attack.

(3) Optimization of anonymous set: To resist the background attack, we established an optimization model based on the Stackelberg Game between users and attackers of background attack to obtain the optimal protection strategy. According to this protection strategy, the construction of the anonymous set is optimized and can resist the background attack.
IV. OUR PROPOSED METHOD

In this section, we first present the system architecture of our proposed method. Secondly, we show the completion method of missed location and construct the candidate location set. Thirdly, we present the construction method of anonymous set under differential privacy protection. Finally, we present the optimization methods of anonymous set to against user background attacks.

A. SYSTEM ARCHITECTURE

To realize fast LBS privacy protection service, we proposed an LBS anonymous service based on edge computing according to paper [27]. The generation of the anonymous collection is accomplished through the cooperation between the anonymous server (edge server) and the mobile terminal. Based on this architecture as shown in Figure.1, the LBS service flows under privacy protection as follows:

1. Users send service requests and current location to anonymous server by mobile terminal;
2. The anonymous server received the user’s request and generates the anonymous set according to proposed method.
3. In order to improve the generation efficiency of anonymous sets, the anonymous set generation in our method is decomposed into several subtasks.
4. Then these subtasks migrate to multiple edge servers mobile terminals and cloudy server for execution based on the task decomposition function in edge computing.
5. After subtask execution completed, the anonymous server aggregate results of subtasks to generate anonymous set and send to LBS server to query.
6. The LBS server sends the query result of anonymous set which include anonymous location and true request location to the anonymous server.
7. The anonymous server filters anonymous result set according to the user’s true location and replies to users’ queries.

B. CONSTRUCTION OF ANONYMOUS CANDIDATE SET

First of all, we use a $m \times n$ dimension matrix $X$ to represent the number of $m$ user’s request location at $n$ different times, which collected by the anonymous server. where $x_{ij} \in X$, $0 \leq i \leq m$, $0 \leq j \leq n$, $m \times n \geq 2k$ represents the location of the user $m$ at time $n$.

Due to the above mentioned reason such as machine failure, there are some missing data in the $X$. In order to reflect if the location is missed in the anonymous server, we give a measure matrix $M$.

$$M = X \times Q, \quad (1)$$

where the element $m_{ij}$ in $M$ reflects whether the location of user $i$ at time $j$ is complete and $\times$ represents an operator of dot product. The matrix $Q$ represents an indicator matrix where $q_{ij} \in Q$ and

$$q_{ij} = \begin{cases} 
1, & \text{if no data for user } i \text{ at time } j \\
0, & \text{otherwise} 
\end{cases}, \quad (2)$$

In order to improve anonymous success ratio, the first task of constructing anonymous candidate set is to complete the missing location in the anonymous server. As the correlation and similarity among users movement trajectories, the matrix $X$ is a low-rank matrix [15]. Then, we can obtain $\bar{X}$ including the complete location of all users by the compression sensing technology [16]. That means, the completion of missing location in $X$ can be converted to solve $\bar{X}$, which satisfies the problem of $\|X - \bar{X}\|_F$ minimization, where $\| \cdot \|_F$ is an operator of the Frobenius Norm.

It is impossible to directly apply aforementioned method to get $\bar{X}$ as the original matrix and the proper rank is unknown. As a good completion $\bar{X}$, it is reasonable to be as close to the matrix $M$ as possible. Thus, we can find the low rank completion by following equation (3)

$$\min_{\bar{X}} \text{rank}(\bar{X}) \quad (3)$$

$$Q \times \bar{X} = M \quad (4)$$

Because equation (3) has the characteristic of non-convex function, it is difficult to directly solve its solution as it is an NP-Hard problem. Since the kernel norm of a matrix is the most compact convex envelope of the matrix rank [17], the solution of the above non-convex function can be transformed into solving convex function as follow:

$$\min_{\bar{X}} |\bar{X}|_F$$

$$Q \times \bar{X} = M \quad (4)$$

In order to find the $\bar{X}$ that satisfies equation (4), we can decompose it based on the matrix decomposition technique [18], which is equivalent to

$$\bar{X} = LR = X_1X_2^T \quad (5)$$
where $X_1 = L^T \tilde{X}$ and $X_2 = R^T \tilde{X}$. Based on various possible decomposition results by equation (5), the optimal is $X_1$ and $X_2$ should minimize the norm of $F$. That is $X_1$ and $X_2$ satisfy formula (6) as follow

$$
\min \|X_1\|_F^2 + \|X_2\|_F^2
$$

$$
Q \times X_1X_2^T = M
$$

(6)

When we obtain $X_1$ and $X_2$ by the strictly satisfy the constraint in equation (6), the solution may be fail for the following reasons. First, The over-fitting problem happened due to there are noises in the location data and strictly satisfaction. Second, the rank of $\tilde{X}$ is only approximate an low-rank matrix. Based on the above analysis, we can convert equation (6) into a convex optimization problem of equation (7) based on Lagrange Method.

$$
\min \left[ (Q \times X_1X_2^T - M) + \lambda \left( \|X_1\|_F^2 + \|X_2\|_F^2 \right) \right]
$$

(7)

To solve the optimization problem in equation (7), we refer literature [19] and propose an iterative algorithm to solve the problem. As shown in algorithm line 7 to line 15. It starts with a random initialization with matrix $X_1$. Then it fixes matrix $X_1$ and solve for matrix $X_2$ by the least square method. Next, $X_2$ is fixed and obtain $X_1$ by the least square method. After a fixed number of iteration $k_1$, we can acquire the optimal completion matrix.

After obtain the complete user location around request location, we select the anonymous candidate location from these satisfies the realistic constraint on two aspects. On one hand, the candidate location should satisfy the Nearest Neighbor Constraint (NNC) based on minimum anonymous area $A_{\min}$ that to against the adversary can infer request location as its anonymous location distributed too concentrated. It requires the distance between the request location and candidate location should be larger than NNC.

$$
NNC = \sqrt{\frac{A_{\min}}{\pi}}
$$

(8)

On the other hand, it should satisfy Maximum Reachability Constraint (MRC), which aims to limit on the maximum distance that a user can move from the request location to the next at fixed intervals. this require the distance to be less than MRC and can be computed as

$$
MRC = \bar{v_{\text{his}}} \times (t_j - t_{j-1})
$$

(9)

where $(t_j - t_{j-1})$ represent the time interval from the previous moment $t_{j-1}$ to the current moment $t_j$ and $\bar{v_{\text{his}}}$ is the request user’s historical average velocity of user. Based on the above two constrains, we select the number of $2k$ desired locations from X to join the anonymous candidate set. Meanwhile, an anonymous candidate set ($AC$) generation algorithm is introduced based on the above steps and is shown in algorithm 1.

**Algorithm 1 Anonymous Candidate Set Construction**

**Input:**
1. The matrix of multi-user original location, $X$;
2. The evaluation matrix $Q$;
3. The weight of Lagrange method, $\lambda$;
4. The parameters, $2k$;

**Output:**
5. The Anonymous candidate set, $AC$;
6. Initialize the $F'$;
7. $M = X \times Q$;
8. for $i = 1$ to $k_1$ do
9. \hspace{1em} $X_1 = \text{get}(L, Q, M)$; // obtain the decomposition matrix;
10. \hspace{1em} $X_2 = \text{get}(L, Q, M)$; // obtain the decomposition matrix;
11. $F = (Q \times X_1X_2^T - M) + \lambda (\|X_1\|_F^2 + \|X_2\|_F^2)$;
12. if $F < F'$ then
13. \hspace{1em} $X_1 = X_1$, $X_2 = X_2$, $F' = F$;
14. end if
15. end for
16. $\tilde{X} = X_1 \times \bar{x}$;
17. $AC = \emptyset$;
18. caculate MRC by formlate(7);
19. caculate NNC by formalte(8);
20. for $i = 1$ to $m$ do
21. \hspace{1em} for $j = 1$ to $n$ do
22. \hspace{2em} $d_{i,j} = \sqrt{(x_{1,i} - x_{1,j})^2 + (x_{2,i} - x_{2,j})^2}$;
23. \hspace{1em} //calculate the distance between real request location and collected $x_{i,j}$ by longitude(lng) and latitude(lat);
24. \hspace{1em} if $d_{i,j} < MRC$ and $d_{i,j} >= NNC$ then
25. \hspace{2em} $AC = AC \cup x_{i,j}$;
26. \hspace{2em} count = count + 1;
27. \hspace{1em} end if
28. \hspace{1em} if count == $2k$ then
29. \hspace{2em} break;
30. \hspace{1em} end if
31. end for
32. end for
33. return $AC$;

**C. CONSTRUCTION OF ANONYMOUS SET**

In anonymous set construction, the key point is the selection of anonymous location. To protect privacy of the location selection, Differential Privacy is introduced. The Differential Privacy is originated from statistical database. It can provide strong protection for individual data when publishing the aggregate information [20], [21], [30]. In this paper, given a location selection algorithm SA and range(SA) represent the range of results selected. For a candidate anonymous set $AC$ and its adjacent candidate anonymous set $AC'$ that differing by only a single location data, the location selection standard differential privacy is defined as follow:

$$
\frac{\Pr(\text{SA}(AC) = l)}{\Pr(\text{SA} (AC') = l)} \leq \exp(\varepsilon)
$$

(10)

where $l \in \text{range}(SA)$ and $\varepsilon$ is the parameters of privacy budget.
The implementation mechanism of differential privacy includes the exponential mechanism and noise mechanism. Instead of directly selecting a location, we use the Exponential Mechanism in Differential Privacy to select anonymous location. As the differential privacy mechanism against attacks in any context and the exponential mechanism can prevent the selection of location from being stolen during the data selection process, location selection can be prevented selection probability leakage by using it. Literature [22] first proposed an exponential mechanism, which mainly deals with some algorithms whose output results are non-numeric, such as the selection of splitting attributes in classification operations and histogram selection in histogram data aggregation published. The key to this mechanism is how to design the scoring function \( u(AC, AC_i) \), where \( AC_i \) represents the selected location from \( AC \). The exponential mechanism that satisfies differential privacy is represented as follows:

\[
A(AC, u) = \left\{ AC : |\Pr[AC \in \text{range}(SA)] \right\} \in \exp \left( \frac{\epsilon u(AC, A \ CAP)}{2\Delta u} \right) \tag{11}
\]

where \( \Delta u \) is the global sensitivity of the scoring function. According to Equation (11), the higher the score \( u(AC, AC_i) \) of location \( AC_i \) is, the higher the probability of \( AC_i \) being selected to join the anonymous set.

Despite the exponential mechanism that was applied can protect privacy in this process, the anonymous location still be identified by the probabilistic attack. The probabilistic attack means the adversary can identify the current real request location according to unevenly distributed request probabilistic at a single-moment and unevenly distributed anonymous in different moment. The single-moment attack means that the attacker filters out the corresponding anonymous location based on the difference between historical request probability of anonymous location and the historical request probability of real location. The continuous-moment attack means that the probability of distribution in the current anonymous set is differs from the probability of historical distribution and the attacker can infer the corresponding position with this knowledge. The measure of probability of historical distribution such as anonymous set entropy [23].

To preserve privacy under the probabilistic attack, the exponential mechanism scoring is design based on above analysis and include two parts of resist single - moment attack and resist continuous - moment attack. In order to resist the single-moment attack, the anonymous locations should be as equal to the historical request probability of current request location as possible. To satisfy this constraint and make it more convenient to measured, we propose a Location Privacy Metric (LPM) to the position probability distribution in the anonymous set based on the entropy. It can be computed as

\[
H = - \sum_{i=1}^{k} p_i \log_2 p_i \tag{12}
\]

where \( p_i \) represents the historical request probability of location \( AC_i \) in anonymous set and

\[
p_i = \frac{\text{number of historical request in location } i}{\text{number of historical request in all locations}} \tag{13}
\]

We know that the max location privacy metric means that the privacy level is the highest and the attacker can’t steal the real location. It can be achieved when the location in the anonymous set has the same history request probability with current request location probability and the LPM will be \( H_{\text{max}} = \log_2 p \). Thus, We use \( H_{i} \) to represent location after \( AC_i \) joined anonymous set according to the resist single-attack part of scoring function design to achieve max location privacy metric as follow:

\[
u_1(AC, A \ CAP) = -|H_{\text{max}} - H_{i}| \tag{14}
\]

According to equation (13), the smaller the difference between \( H_{i} \) and \( H_{\text{max}} \), the easier \( AC_i \) is to be selected because of its stronger ability to resist the attack. To resist the continuous-time attack, the selected location should satisfy the constraint that the LPM of its current anonymous set distribution is equal to the historical anonymous set distribution. The historical anonymous set represent the anonymous sets whose constructed before current time. In order to satisfy this constraint, we propose an anonymous privacy metric based on variance. As we know, it is difficult to directly apply variance to the above constrain as its huge computational overhead. Based on the above reasons, we use the average entropy variance of the anonymous set. We first compute the average entropy of each historical anonymous sets as

\[
\bar{H} = -\frac{\sum_{i=1}^{n} H_{i}}{n} \tag{15}
\]

Then, the average variance of historical anonymous set can be obtained by

\[
\sigma^2 = E \left[ (H - \bar{H})^2 \right] = -\frac{\sum_{i=1}^{h} (H_{i} - \bar{H})^2}{h} \tag{16}
\]

where \( h \) represents the number of historical anonymous set.

After the selected location \( AC_i \) joins the anonymous set, the smaller the average variance difference between the anonymous set and historical anonymous sets is, the smaller the difference between it and the historical anonymous set is, and it has a stronger ability to resist probability attack. In this way, we can design the resist continue-moment attack part of score function based on current anonymous average variance entropy after \( AC_i \) joins as

\[
u_2(AC, A \ CAP) = -\frac{|\sigma^2 \text{ variance} - \sigma_i^2 \text{ variance}|}{\sigma^2 \text{ variance}} \tag{17}
\]

A smaller the difference between \( \sigma^2 \) and \( \sigma_i^2 \) indicate that \( AC_i \) has more stronger ability to resist attacks and can be selected. Combine the above two scoring functions, the scoring function \( u(AC, AC_i) \) of location selection probability can be obtained show in equation(18) and the laplace noise is
added to the anonymous set after complete location selection according to the [24].

\[
s_i = \Pr(AC_i) = \frac{\exp\left(\frac{\epsilon u(AC_i|C_i)}{2\Delta n}\right)}{\sum_{AC_j \in AC} \exp\left(\frac{\epsilon u(AC_j|C_j)}{2\Delta n}\right)}
\]  

(18)

As the value range of the equation (15) and the equation (16) is [0,1], so \(\Delta n = 1\). Based on the results of the descending formula, we obtain the sequence of location and selected the first \(k\) location to construct the anonymous set (AD) as shown in algorithm 2.

**Algorithm 2 Anonymous Set Construction**

**Input:**
1. The Anonymous candidate set, AC;
2. The Differential privacy budget, \(\epsilon\);
3. The iterations, \(k\);
4. The privacy parameters, \(k\);

**Output:**
5. The Anonymous set, AD;
6. \(C = \emptyset\);
7. \(S = \emptyset\);
8. \(\epsilon_1 = 0.5 \times \epsilon\);
9. \(\epsilon_2 = 0.5 \times \epsilon\);
10. calculate the \(\delta^2\) according to the formula (15);
11. calculate the \(H_{\text{max}}\) according to the formula (14);
12. for \(i = 1\) to \(|AC|\) do
13. calculate the \(H_i\) according to the formalte (12);
14. \(\mu'_i = H_{\text{max}} - H_i\);
15. calculate the \((\delta_i)^2\) according to the formalte (15);
16. \(\mu'_2 = \delta^2 - \delta_i^2\);
17. \(\mu' = \mu'_i + \mu'_2\);
18. Combine \(\mu'\), \(\epsilon_1\) with formula (17) to get the selected probability \(s_i\) of \(AC_i\);
19. \(S = S \cup s_i\);
20. end for
21. for \(i = 1\) to \(k\) do
22. if \(s_i == \text{MAX}(S)\) then
23. \(AC_i = AC_i + \text{lap}(\epsilon_2)\);
24. \(AD = AD \cup AC_i\);
25. \(S = S - s_i\);
26. end if
27. end for
28. return AD

**D. OPTIMIZATION OF ANONYMOUS SET**

The k-anonymity can resist traditional attacks to some extent, but it cannot prevent background attacks. Background with knowledge such as privacy protection mechanism etc, the attacker can infer the request location in the anonymous set.

For the anonymous collection AD constructed in section C, the background knowledge includes the initial distribution probability of anonymous data and the anonymous protection method. Based on this background knowledge, the objective of adversary is to build an optimized attack strategy \(Q'\) that can infer the location \(AD'_i\) as close to the request location \(AD_i\) as possible. The attack probability of adversary as

\[
Pr(AD_i|AD) = \frac{Pr(AD, A D_i)}{Pr(AD)} = \frac{Pr(AD)(AD_i)(\varphi(AD_i))}{\sum_{AD_i \in AD} Pr(AD)(AD_i)(\varphi(AD_i))}
\]  

(19)

The protector take this strategy as a parameter input to the optimization model to optimize the construction of anonymous collections. The ultimate purpose of protector is to make the location inferred by the attacker as far away from the real location as possible.

In order to realize the purpose of the protector, we can build an optimization model based on Stackelberg Game model. In the optimization model, the protector is reckoned as leader, and the attacker as follower. The game requires the leader to take strategy first and then the follower. The followers execute their strategies based on the observed strategy of the leader. Thus, after protector execute their strategies and give a anonymous set, the adversary optimize his attack strategy based on the anonymous set. Since this game is a leader-first model, its result under the Nash Equilibrium is the optimization protection strategy \(P'\) against the optimization attack. In other words, the optimization of AD can be accomplished based on this mechanism \(P'\).

In this game model, the benefit of participants is the distance \(d(AD_i, AD'_i)\) between the real request location \(AD_i\) and the inferred location \(AD'_i\). A greater \(d(AD_i, AD'_i)\) means a greater benefit of protector and a higher protection level of the anonymous collection. A smaller \(d(AD_i, AD'_i)\) denotes a greater benefit of the attacker and a lower protection level of the anonymous collection and thus AD needs to be optimized.

In order to realize the optimization of protection strategy, the protector first use the anonymous collection AD generated in the previous section as his initial strategy. After the attacker makes inferential attacks strategy based on AD, the attack benefit expectations can be expressed as

\[
\sum_{AD} Q(AD'|AD) d(AD'_i, A D_i)
\]  

(20)

Then, the protector adjusts his protection strategy according to above adversary’s attack strategy to maximize his protection benefit. Then his protection benefit expectations can be expressed as

\[
P(AD) \sum_{AD} Q(AD'|AD) P(AD|AD_i) d(AD'_i, A D_i)
\]  

(21)

where \(P(AD) = \sum_{AD} \varphi(AD)P(AD|AD_i)\) as the probability of anonymous collection AD.

The optimized attack strategy is one that minimizes the above expectations and as follow

\[
\min_{AD} \sum_{AD} Q(AD'|AD) d(AD'_i, A D_i)
\]  

(22)

The optimized protect strategy is the one that maximize the above expectations gainst the optimized attack strategy.
Based on equation (22), it is computed as
\[
\arg \max_{AD} \min_{AD} \sum_{AD} P(AD|AD_i) Q(AD_i|AD) d(AD_i, A D_i)
\]
(23)

In this Stackelberg Game, the adversary aims to obtain the attack strategy that maximizes the attack benefit, and the protector base and the attack strategy obtain the protection strategy that maximizes the protection benefit.

The purpose of Stackelberg Game is to find the Strong Nash Equilibrium (SSE) that the adversary unable to use the optimized attack strategy to gain more benefits. In other words, it makes the adversary can't infer the real request position. In this game, SSE chooses the most beneficial strategy for the leader (protector) when the adversary gains the same benefits from multiple strategies. Therefore, the Nash Equilibrium can be obtained and the optimal protection strategy can be solved under this equilibrium. Combining with (22)(23), we can use (24) to obtain optimal protection \( AD' \) under the aforementioned Nash Equilibrium
\[
AD' = \arg \max_{AD} P(AD) \min_{AD} \sum_{AD} P(AD|AD_i) Q(AD_i|AD) d(AD_i, A D_i)
\]
(24)

Based on the above formulas and the availability constraints \( \beta \), we use the linear programming model to obtain the optimal protection strategy

- \( \max P(AD) \sum_{AD} P(AD|AD_i) Q(AD_i|AD) d(AD_i, A D_i) \)
- \( \sum_{AD,AD'} \phi(AD_i) P(AD|AD_i) d(AD_i, A D_i) \leq \beta \)
- \( P(AD) \sum_{AD} Q(AD_i|AD) d(AD_i, A D_i) \)
- \( s.t. \sum_{AD} Q(AD_i|AD) = 1 \)
- \( Q(AD_i|AD) \geq 0 \) (25)

To solve the optimal \( AD' \) based on Equation (25), the optimal attack strategy \( Q' \) needs to be obtained in advance based on [28], which is difficult to solve. In order to simplify the solving process, we introduce the minmax method to solve the problem. In this solving, the max-min strategy of the protector can be solved to be the optimal protection strategy while minimizing the maximum profit of the adversary and equal to

\[
\max \sum_{AD,AD_i} \phi(AD_i) Q(AD_i|AD) P(AD|AD_i) d(AD_i, A D_i) \geq z
\]
\[
\sum_{AD,AD_i} \phi(AD_i) P(AD|AD_i) d(AD_i, A D_i) \leq \beta
\]
\( s.t. \sum_{AD} Q(AD_i|AD) = 1 \)
\( Q(AD_i|AD) \geq 0 \) (26)

where \( z = \min \sum_{AD_i \in AD} \phi(AD_i) Q(AD|AD_i) d(AD_i, A D_i) \)

E. SECURITY ANALYSIS

From the description of our method, we can see that the algorithm guarantees the security from two main steps:

In the construction of anonymous set, the algorithm 2 adopts the exponential mechanism to select anonymous data and adds Laplace Noise to it with the privacy budget. According to the combination theorem of the differential privacy [12], algorithm 2 realizes differential privacy. With the powerful protection capability of differential privacy, algorithm 2 can effectively prevent attackers from inferring the probability of anonymous data selection, thus prevent the privacy leakage problem of anonymous set construction. At the same time, due to the reasonable design scoring function of the exponential mechanism for data selection, the probability of the selected data meeting the historical request is very close, and the data distribution of the anonymous set in different time is very similar to that of the historical anonymous set, thus effectively prevent privacy leakage under the probabilistic attack.

After completing the construction of the anonymous set. The anonymous set in this paper include number of \( k - 1 \) others location make the real request location disclosure risk less than \( k \) even if the attacker knows the location distribution. In additional, the different differential privacy in construct anonymous set. Moreover, the Stackelberg Game model is established according to (26) to optimize the anonymous data which increases the difficulty of background attack. Combining above steps, our method is secure and can effectively protect the user’s location information in above attack.

V. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

A. EXPERIMENTAL SETTING

We evaluate the performance of the proposed privacy-preserving method based on Geolife Dataset, which collected 5 years of trajectory data from 182 users and contains 17621 trajectories in Beijing. As there is no user query service in Geolife Dataset, we simulate to generate the query for each location under the real conditions and the content is replaced by numbers. All experiments have been repeated more than 200 times and each measurement is averaged over 30 instances. The final result comes from the average of this multiple experiments. The \( k \) is commonly set from 3 to 5, \( mp = 5\% \), \( b = 0.1 \), where \( mp \) represent the proportion of missing data in complete data and it is realized by random sampling, \( b \) represent the background knowledge.

We apply 4 metrics to evaluate the performance of our method. In order to evaluate the privacy protection effect of our method, the privacy level (PL) is introduced and it is computed by formulait(20). To verify the time efficiency of the proposed method, the Response Time (RT) is used for evaluation [29]. In order to evaluate the availability of our algorithm, the location anonymous area (AA) is used and it computed by literature [9]. To evaluate the privacy protection capability of our method under the attack, we use...
the Probability of Exposing Real Location (PERL) and it is computed as
\[
\text{PERL} = \frac{1}{k - k' + \tau}
\]  
(27)
where \(k'\) represent the adversary has inferred anonymous location and \(\tau\) represents the minimum constraint factor and the denominator is equal to 0.

**B. EFFECT OF PRIVACY PROTECTION**

We first verify our method for privacy protection in the missing anonymous candidate locations (It is equivalent to incomplete collection set). As can be seen from figure 2a, the AA curve of our method tends to increased with the value of \(mp\) increases and grow slower after \(mp = 2.50\%\) under different \(k\) also a higher \(k\) need larger AA. The figure 2b shows the AA curve with same tendency under different \(mp\). The higher AA represents Anonymous cost and low privacy protection in the missing anonymous candidate locations and vice versa. The AA curve increased reason that it is not possible to find further anonymous locations to build anonymous collections as the number of missing locations increases. The slower growth indicated that the result of completion missing location in our method completed the missing location through the relationship between multiple users. In additional, this complement more and more obviously with the increase of missing location. That means it is very effective to protect the privacy in the missing location according to our method.

We secondly verify our method for privacy protection in the anonymous collection construct process. As we can seen from figure 3a, the PERL curve has a decreasing trend with the increase of privacy budget and a higher \(k\) has a lower probability of exposure. Figure 3b tell us the PERL curve has a decreasing trend with the value of privacy budget increase and a higher \(mp\) made the anonymous location easier to be identified. The reason is that the increase of privacy budget makes there are more budget choices to meet the privacy protect requirement of anonymity and against the probability attack. Since differential privacy based on privacy budget can prevent privacy disclosure in anonymous process, it also indicates...
that there is more powerful privacy protection capabilities in this process. Based on the above analysis, our method can alleviate the privacy leakage problem in anonymous set construction.

We thirdly verify our method for privacy protection against the background attack. We use the equations (18) to measure the adversary background knowledge. As can be seen from figure 4, the increasing tend of PL curve is become more obvious with the increased $b$ and indicates that the proposed method can effectively against the background attacks. This is due to the fact that we use the game model to optimize the protection strategy and obtain more significant gains. The figure 4 tell us The larger the $k$ value, the better effect of the anonymity protection can be.

We finally verify time efficiency of our method for privacy protection. As can be seen from figure 5, the increase tend of RT curve is becoming the trend is slowing down with the increased $k$ in different mp and. It indicates that our method can provides time-feasible privacy protection. This is because our approach combines privacy optimization after simplified and effectively data selection.

C. EFFECT COMPARISON OF DIFFERENT ALGORITHMS

In this section, we evaluate the privacy protection effect based on the PL, AA, PERL with CirDummy [8] and DLS [9]. Because above methods lack the completion of missing position, we complete the missing data by Kalman Filtering method [26]. As can be seen from Figure 6 with the increase
of $k$ value, the AA of all the three algorithms showed an upward trend in general. Compared with CirDummy and DLS algorithm, our algorithm increases slower because of its completion of the surrounding missing location method and it can select more close location to be anonymous. As the restriction of the anonymous region of CirDummy interacts with its discreteness requirement, its AA is worse than our algorithm which lead to a rapid decrease of query accuracy lower than our algorithm.

Figure.7 is the comparison of PL of three algorithms with the change of the number of $k$. As can be seen from the figure, with the increase of the number of $k$, there is an increasing trend in PL of these algorithm. Due to our method considering the anonymous historical request probability in the selection of anonymous data, it has the fastest growth. Because the GridDUMMY’s dummy location has an equal historical probability in the selection of anonymous data, GridDUMMY’s PL is better than DLS.

As can be seen from figure 8, the PERL of three algorithms tends to decline as the value of $k$ increases, which means that the higher the value of $k$ is set, the better the privacy protection effect of the algorithm is. Meanwhile, because our algorithm adopts the method of attack optimization, it achieves best result than the other two methods. That means our method has the best ability to resist background attack.

VI. CONCLUSION

In this paper, we present a privacy-protection scheme to protect user’s request location in LBS based on k-anonymity. In order to solve the problem of low anonymous success rate, the anonymous candidate set establishment method for incomplete collection location reconstruction is proposed. To prevent the disclosure of the selection probability in the process of anonymous set construction, the anonymous set is constructed based on the differential privacy method. Finally, we establish the background knowledge of Stackberg Game model to resist the attacker and the experiment proves that our method is very effective. On this basis, we will study how to ensure the quality of the anonymous set to satisfy user privacy preferences when the anonymous candidate location is missing in the future work.

REFERENCES

[1] A. Khoshogzaran, C. Shahabi, and H. Shirani-Mehr, “Location privacy: Going beyond K-anonymity, cloaking and anonymizers,” Knowl. Inf. Syst., vol. 26, no. 3, pp. 435–465, Mar. 2011.
[2] G. Sun, S. Cai, H. Yu, S. Maharanj, V. Chang, X. Du, and M. Guizani, “Location privacy preservation for mobile users in location-based services,” IEEE Access, vol. 7, pp. 87425–87438, 2019.
[3] T. T. B. Le and T. K. Dang, “Semantic-aware obfuscation for location privacy at database level,” in Proc. Inf. Commun. Technol.-EurAsia Conf. Berlin, Germany: Springer, 2013, pp. 111–120.
[4] M. Han, L. Li, Y. Xie, J. Wang, Z. Duan, J. Li, and M. Yan, “Cognitive approach for location privacy protection,” IEEE Access, vol. 6, pp. 13466–13477, 2018.
[5] M. L. Yiu, C. S. Jensen, J. Möller, and H. Lu, “Design and analysis of a ranking approach to private location-based services,” ACM Trans. Database Syst., vol. 36, no. 2, p. 10, May 2011.
[6] P. Perazzo and G. Dini, “A uniformity-based approach to location privacy,” Comput. Commun., vol. 64, pp. 21–32, Jun. 2015.
[7] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, “Quantifying differential privacy in continuous data release under temporal correlations,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 7, pp. 1281–1295, Jul. 2019.
[8] H. Lu, C. S. Jensen, and M. L. Yiu, “PAD: Privacy-area aware, dummy-based location privacy in mobile services,” in Proc. 7th ACM Int. Workshop Data Eng. Wireless Mobile Access, 2008, pp. 16–23.
[9] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, “Achieving K-anonymity in privacy-aware location-based services,” in Proc. IEEE Conf. Comput. Commun. (INFOCOM), Apr. 2014, pp. 754–762.
[10] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, “Enhancing privacy through caching in location-based services,” in Proc. IEEE Conf. Comput. Commun. (INFOCOM), Apr. 2015, pp. 1017–1025.
[11] X. Zheng, Z. Cai, J. Li, and H. Gao, “Location-privacy-aware review publication mechanism for local business service systems,” in Proc. IEEE Conf. Comput. Commun. (INFOCOM), May 2017, pp. 1–9.
[12] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, “Preventing location-based identity inference in anonymous spatial queries,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 12, pp. 1719–1733, Dec. 2007.
[13] T. Xu and Y. Cai, “Feeling-based location privacy protection for location-based services,” in Proc. 16th ACM Conf. Comput. Commun. Secur. (CCS), 2009, pp. 348–357.
[14] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, “Limits of location privacy under anonymization and obfuscation,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2017, pp. 764–767.
[15] W. Zhao, B. Sun, T. Wu, and Z. Yang, “On-chip neural data compression based on compressed sensing with sparse sensing matrices,” IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 1, pp. 242–254, Feb. 2018.
[16] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, “A compressive sensing approach to urban traffic estimation with probe vehicles,” IEEE Trans. Mobile Comput., vol. 12, no. 11, pp. 2289–2302, Nov. 2013.
[17] L. Kong, L. He, X.-Y. Liu, Y. Gu, M.-Y. Wu, and X. Liu, “Privacy-preserving compressive sensing for crowdsensing based trajectory recovery,” in Proc. IEEE 35th Int. Conf. Distrib. Comput. Syst., Jun. 2015, pp. 31–40.
[18] Z. Yu, C. Gu, Z. Jing, Q. Cai, Y. Luo, and Y. Wang, “Cryptanalysis of an asymmetric cipher protocol using a matrix decomposition problem: Revisited,” Multimedia Tools Appl., vol. 77, no. 9, pp. 11307–11320, May 2018.
[19] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, “Quantifying differential privacy under temporal correlations,” in Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE), Apr. 2017, pp. 821–832.

[20] M. Ye and A. Barg, “Optimal schemes for discrete distribution estimation under locally differential privacy,” IEEE Trans. Inf. Theory, vol. 64, no. 8, pp. 5662–5676, Aug. 2018.

[21] L. Wang, D. Yang, X. Han, T. Wang, D. Zhang, and X. Ma, “Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation,” in Proc. 26th Int. Conf. World Wide Web, Apr. 2017, pp. 627–636.

[22] G. Brandli and M. Dick, “Alternating current fed power supply,” U.S. Patent 4,084,217, Nov. 4, 1978.

[23] C. Dwork, “Differential privacy: A survey of results,” in Proc. 5th Int. Conf. Theory Appl. Models Comput. (TAMC), 2008, pp. 1–19.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith “Calibrating noise to sensitivity in private data analysis,” in Proc. Theory Cryptogr. Conf. Berlin, Germany: Springer, 2006.

[25] S. V. Kumar, “Traffic flow prediction using Kalman filtering technique,” Procedia Eng., vol. 187, pp. 582–587, Jan. 2017.

[26] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.

[27] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, “Protecting location privacy: Optimal strategy against localization attacks,” in Proc. ACM Conf. Comput. Commun. Secur., 2012, pp. 617–627.

[28] R. Shokri, “Optimal user-centric data obfuscation,” Proc. Privacy Enhanc- ing Technol., vol. 2015, no. 2, pp. 299–315, 2014.

[29] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and P. S. Yu, “LoPub: High-dimensional crowdsourced data publication with local differential privacy,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.

[30] R. Gupta and U. P. Rao, “Achieving location privacy through CAST in location based services,” J. Commun. Netw., vol. 19, no. 3, pp. 239–249, 2017.

[31] H. Kido, Y. Yanagisawa, and T. Satoh, “An anonymous communication technique using dummies for location-based services,” in Proc. Int. Conf. Per- vas. Services (ICPS), 2005, pp. 88–97.

[32] R. Gupta and U. P. Rao, “An exploration to location based service and its privacy preserving techniques: A survey,” Wireless Pers. Commun., vol. 96, no. 2, pp. 1973–2007, Sep. 2017.

XUDONG YANG received the M.S. degree from Northwest University, China, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include privacy protection and edge computing.

LING GAO received the Ph.D. degree in computer science from Xi’an Jiaotong University, Xi’an, in 1996. He is currently a Professor with the Department of Computer Science, Northwest University. His research was sponsored by several National Science Foundations. His research interests include privacy protection and edge computing.

JIE ZHENG received the Ph.D. degree in computer science from Xi’an Dianzi University, Xi’an, in 2008. He is currently an Associate Professor with the Department of Computer Science, Northwest University. His research was sponsored by several the National Science Foundations. His research interests include privacy protection and edge computing.

WEI WEI (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Xian Jiaotong University, in 2005 and 2011, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Xi’an University of Technology, Xi’an, China. He ran many funded research projects as a Principal Investigator and Technical Members. He has published around 100 research articles in international conferences and journals. His research interests are in the area of wireless networks, wireless sensor networks application, image processing, mobile computing, distributed computing, pervasive computing, the Internet of Things, sensor data clouds, and so on. He is an Editorial Board Member of FGCS, AHSWN, IEICE, KSII, and so on. He is a TPC Member of many conferences and regular Reviewer of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS (TPDS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON MOBILE COMPUTING (TMC), the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (TWC), and many other Elsevier journals.

95854 VOLUME 8, 2020