Computing (2022) 104:1739–1751
https://doi.org/10.1007/s00607-022-01067-4

REGULAR PAPER

All-dummy k-anonymous privacy protection algorithm based on location offset

Jianghui Liu[1] · **Shengxiang Wang**[1]

Received: 24 May 2021 / Accepted: 16 February 2022 / Published online: 13 March 2022

Abstract
While k-anonymous algorithms can effectively protect users' private location information, the problem of selecting an appropriate location in the anonymous area to construct the k-anonymous area remains a significant one. When selecting real users from the surrounding area to co-construct anonymous regions, it is easy to cause the leakage of user location information. Moreover, using false addresses to construct a region requires calculating the probability of location queries, which increases the computational complexity. In this paper, an all-dummy k-anonymous algorithm based on location offset is proposed to construct anonymous regions. This algorithm randomly selects k−1 locations and real users in the selected anonymous compose an anonymous group at first. Subsequently, these coordinates are centered on migration, generating multiple dummy addresses of each location migration, such that the dummy address distance is greater than the radius of the user's query, with the dummy address location information used for the location server queries. Through experimental verification, compared with the circle-based dummy address generation algorithm and the random k-anonymous algorithm, the all-dummy k-anonymous algorithm is found to achieve an entropy value and tracking success rate closer to the optimal k-anonymous algorithm without increasing the communication cost.

Keywords k-anonymous · Location offset · Privacy protection

1 Introduction

While the rapid development of mobile devices has brought great convenience to people's lives, it has also created some problems, one of which is the leakage of

Shengxiang Wang author contributed equally to this work.

✉ Shengxiang Wang
1084019500@qq.com; hl1211_hui@163.com

[1] School of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China

users' private information. Elements of user privacy include the user's location privacy, query privacy and so on. Many algorithms have been proposed in an attempt to solve the problem of location privacy disclosure and consequently protect private location information. The k-anonymous algorithm, a typical algorithm of this kind [1], mixes a user's location together with at least k−1 other user locations to form an anonymous set of k users, thus achieving the purpose of protecting user location privacy.

Although the k-anonymous algorithm can effectively protect users' private location information, it is also affected by a number of shortcomings [2]. First and foremost, the k-anonymous algorithm requires a large amount of computation. Second, it needs to rely on a third-party anonymous server; however, such servers may easily become the targets of attackers, resulting in the disclosure of users' location privacy [3]. The k-anonymous algorithm also needs to use another k−1 users around the original user to collectively form the k-anonymous area. However, studies have shown that if real users are used to form the k-anonymous area, attackers can pose as a user and initiate a query, making it easy to expose the real location of other users. Several studies have accordingly been conducted that use a dummy or cache location to constitute the k-anonymous area. However, the difficulty of this method is the dummy location and cache location choice: if randomly assigned the dummy location, may be assigned to the location of users are unlikely to reach the location, the attacker can rule out the location query, meaning that the degree of privacy protection provided by the k-anonymous algorithm is reduced.

To solve this problem, many solutions have been proposed, the most common of which is to select a dummy location by querying probability. Niu et al. [4] proposed dividing the region into grids, calculating the query probability of all grids, and selecting the grid with a similar probability to the user's real location to generate a dummy location for the query, thereby reducing the probability of an attacker identifying the real user and consequently improving the degree of user privacy protection. However, more computing power is needed to calculate the query probability and choose a similar location in this way. As mentioned in the article, it takes four hours to calculate the points of interest in an 8 km × 8 km map area, which requires a significant amount of computing resources. Niu et al. [5] subsequently proposed improving the protection of user privacy information by caching the user's query information; while this method can reduce the amount of information querying required, the computation is still large. Li et al. [6] proposed coloring the Voronoi polygon by using the four-color mapping theorem to help users select an appropriate virtual location. Tan et al. [7] further proposed that k-anonymity and the semantic trajectory should be combined to form multiple sensitive areas in order to protect user privacy. In addition, some other frameworks have been proposed to solve this problem [8].

The anonymous method based on dummy location can solve the user location selection problem [9]. By generating a dummy location and sending it to the server along with the user's real location for querying, the goal of hiding the user's real location can be achieved. Dummy location generation methods include random dummy location generation and generation of a dummy location under constraints. Comparing the two methods, the random dummy location generation protects the user's private location information to a lesser degree. Generating a dummy location under constraints works to hide the user's real location so as to prevent central and boundary attacks in the

hidden area, making the generated dummy location distribution more uniform and real on the premise that the requirements of the hidden area are met.

While the anonymous method based on dummy location is simple and convenient, the dummy location generated can be easily recognized by the attacker, leading to a further decrease in the degree of privacy protection. In order to make the generated dummy location appear more real, in this article, we propose an all-dummy k-anonymous privacy protection algorithm based on location offset. This algorithm first randomly generates k−1 locations; these locations are then used as locating points to offset location coordinates and generate multiple dummy addresses. The distance between each address is greater than the dummy address query radius, and the algorithm can generate multiple sets of dummy addresses. Using the generated more groups of dummy address query, can protect the user's real location information from attackers. Moreover, because the dummy addresses are generated through the offset of the locating point, the attacker cannot determine that the dummy address is a dummy address even if it is in a location that the user cannot reach; this makes the dummy address appear more real and improves the degree of privacy protection of the user's location information.

2 Related work

In recent years, there are many researches on privacy protection, and various algorithms are used in privacy protection. Yang et al. [10] proposed 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. Lu et al. [11] proposed the PAD approach that is capable of offering privacy-region guarantees. PAD uses so-called dummy locations that are deliberately generated according to either a virtual grid or circle.

2.1 Entropy-based privacy measurement

In this paper, entropy is used to measure the degree of privacy protection provided to users. Information entropy is used to measure the expected value of a random variable: the greater the information entropy of a variable, the more content it contains, the greater the uncertainty and the higher the degree of privacy protection for variable information [12]. Location entropy is similar to information entropy, which can be used to measure the amount of information obtained by an attacker from an anonymous set. The higher the location entropy, the harder it will be for an attacker to identify the user's real location. Suppose that the probability of a user sending a service request at some location i is p_i, then generate n dummy locations [13]; then the calculation formula of the entropy value can accordingly be represented as:

$$H = -\sum_{1}^{n} q_i * \log(q_i) \qquad i = 1, 2 \ldots n \tag{1}$$

where $q_i = p_i / \sum_1^n p_i$; in general, q_i is normalized so that the sum of q_i is 1.

As can be seen from Eq. 1, the maximum entropy value can be obtained when the probability of sending a request to all dummy locations is equal; at this moment, $H_{max} = \log_2 n$. The closer the probability that all the dummy addresses send location service requests, the closer the entropy is to H_{max}.

According to the algorithm proposed in this paper, the formula is further derived, and the new definition of q_i is given as:

$$q_i = \frac{x_i}{\sum_1^n x_i} \qquad i = 1, 2. \ldots n \tag{2}$$

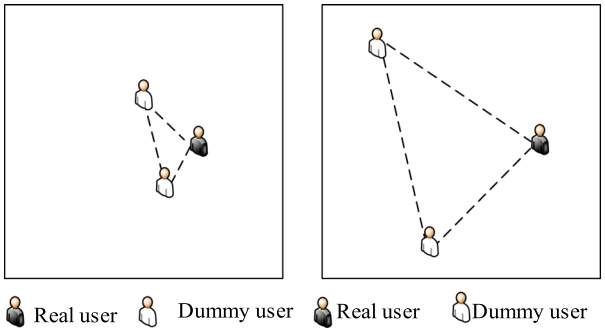
Then,

$$H = -\sum_1^n \frac{x_i}{\sum_1^n x_i} * \log\left(\frac{x_i}{\sum_1^n x_i}\right) \qquad i = 1, 2. \ldots n \tag{3}$$

The location entropy is generally affected by two factors:

The more locations are contained in the anonymous set, the more chaotic the anonymous set will be, and the greater the degree of entropy;
With the same number of locations in the anonymous set, the greater the distance between each location, the more uniform the location distribution in the anonymous space will be, and the higher the entropy value.

As shown in Fig. 1, the user generates two dummy locations to achieve k-anonymity, where k = 3. In Fig. 1a, dummy locations closer to real users are selected, while the dummy locations in Fig. 1b of are far away from real users. In these two cases, the number of location coordinates in the anonymous set is same. When choosing the dummy locations, however, it is more likely that the locations in Fig. 1b will be selected; this is because, at this time, the dummy locations are far away from the user's location, making it more difficult for the attacker to identify the user's real location.



(a) Dummies with smaller distance **(b)** Dummies with bigger distances

Fig. 1 Distribution of dummy locations

Accordingly, when generating the offset location, it is important to design a scheme that causes the offset location to be as far away from the real user as possible.

2.2 Dummy location generation method

The dummy location generation method is one of the methods utilized to protect user location information. This method need generating dummy locations and querying the server together with the real location of the user, meaning that the attacker cannot accurately find the user's real location. Dummy location methods generate dummies with randomly generated locations and constraint conditions. Of the two methods, randomly generated dummy locations will inevitably result in two more bad situations: one is to generate a dummy location in an area that the user may not reach, while the other is a crowd around the dummy location generated for the user. Both of these situations will reduce the degree of privacy protection provided for users' location information. At the same time, the number of dummies should not be too high; too many dummy locations will greatly increase the computing overhead of the location server, resulting in increased waiting time for user service requests.

Methods that generate dummy locations under constraints can be divided into two categories: one is generating dummy locations based on circles, while the other is generating dummy locations based on grids. The former involves setting the anonymous generation region as a circle, dividing the circle region into k equal parts, and taking points on its bisector as dummy locations. Generating dummy locations based on grids involves setting rectangular anonymous regions, dividing them into grids, and taking points on the grids as dummy locations. Generating dummy locations under constraints can avoid the two above-mentioned bad situations to a certain extent while also preventing central and boundary attacks in the hidden area. On the premise that the requirements of the hidden area, the distribution of generated dummy locations is more uniform, while the degree of privacy protection provided for the user's location information is also improved.

In addition, dummy locations are generated via location offset: more specifically, a location near the user is used to query the server rather than the user's actual location, which further improves location privacy protection. In this paper, an improved location offset method is used to generate multiple dummy locations via multiple offsets of user coordinates, improving the degree of user location information privacy protection to a greater extent. The location offset method has the advantage of not increasing the communication overhead when queried. We improved the location of the migration method to achieve similar advantages. The overhead of each dummy location query communication is constant. With the same number of queries, the location of the k-anonymity algorithm results in a comparative increase communication overhead, but also provides a greater degree of protection to the user's private location information.

Table 1 Symbol table

pos	User's real location
R	Anonymous area side length
r	Radius of the query
k	Number of dummy location groups
m	The number of dummy locations in each set
p	Query accuracy
n	Total number of dummy locations

3 All-dummy k-anonymous algorithm based on location offset

3.1 Related definitions and parameters

1. Anonymous area: < S,n > , an anonymous area of area S containing n dummy locations;
2. Anonymous set: $pos^{'} = (pos_1, pos_2, \ldots pos_k)$, a set of locations used to represent a user's query to the server;
3. Query request: $Q = (pos, r, \langle S, n\rangle, req)$, represents query information sent by the user to the server. The query request after using the anonymous algorithm is $Q^{'} = \left(pos^{'}, r, \langle S, n\rangle, req\right)$; here, $pos^{'}$ includes n dummy locations, which can hide the real location information of the user, r is the query radius, n is the number of locations that initiate a query to the server, and req is the content requested by the user.
4. Query result accuracy: Generally refers to the percentage of query results obtained by the location privacy protection method relative to the total query results obtained by using the user's real location request service. Query result accuracy can reflect the influence of location privacy protection method on the query results: the higher the query result accuracy, the closer the simulation is to the real situation and the better the performance of the location privacy protection method.

Table 1 presents the meanings of the symbols used in this article.

3.2 Dummy location generation process

In the all-dummy k-anonymous location privacy protection algorithm based on location offset, the locations generated are all dummy locations; thus, two requirements must be met:

1. After using the algorithm of anonymous sent query request Q', the user's true location must be hidden so as to achieve the aim of protecting users' privacy, while the query results should be as consistent as possible with the results the user obtains through the original query request Q to ensure quality of service is maintained;

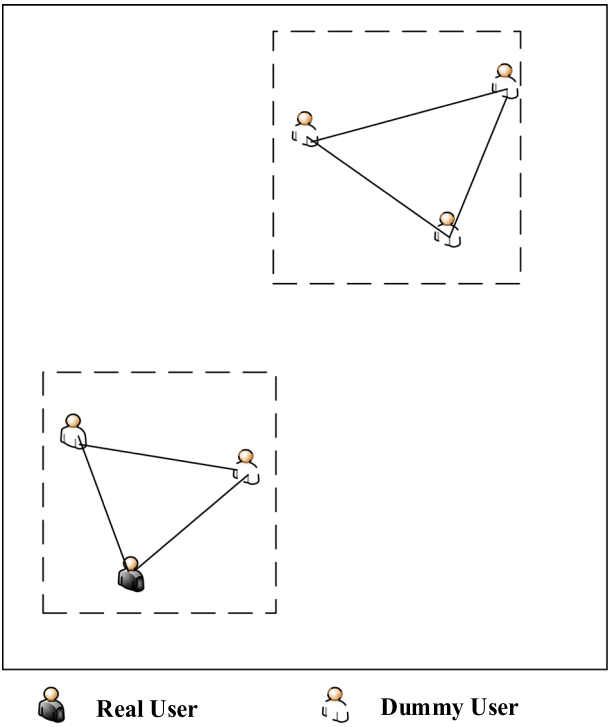


Fig. 2 The dummy location generation process

2. The n generated false addresses must satisfy the condition constraint of the hiding area $< S, n >$.

The process of anonymous user location information processing is illustrated in Fig. 2. In the anonymous area with side length R, k−1 dummy locations are randomly generated, and these k−1 dummy locations are taken as locating points along with the user's real location. Each locating point generates m dummy locations again, and these dummy locations are used for querying. The offset distance depends on the query radius R, and the vertical and horizontal coordinates of the user's real location are offset respectively. The maximum offset distance is R. In order to obtain a larger entropy value, the distance between each offset location and other offset locations in the same group exceeds the query radius R when the offset location is obtained, and the query is carried out from the offset location.

In the process of generating all-dummy locations, the randomly generated dummy locations are dispersed as far from each other as possible, lest a different anonymous query selection group appear in the offset range overlapping phenomenon; moreover, the same set of offset distance is greater than the radius of the query r, making dummy locations' distribution in the anonymous area as even as possible.

The algorithm process of dummy location generation is presented in Table 2.

Table 2 All-dummy k-anonymous algorithm based on location offset

pos	User's real location
R	Anonymous area side length
r	Radius of the query
k	Number of dummy location groups
m	The number of dummy locations in each set
p	Query accuracy
n	Total number of dummy locations

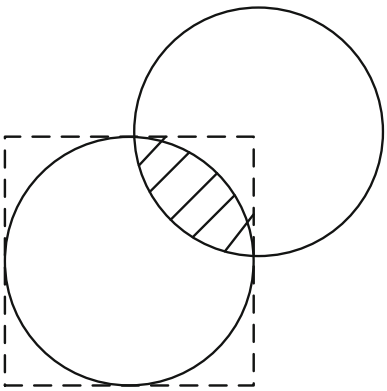


Fig. 3 Offset coordinate location when p is minimum

In this article, the query accuracy $P = S/S_{pos}$, due to the maximum deviation from the user's true location coordinates being r; the more the dummy location deviates from the true location, the smaller the query accuracy. Thus, the minimum value of p in the offset is related to the value of r, as well as the amount of the offset m. When $m = 1$, the minimum value is obtained when the offset location coordinates $pos' = (pos_x + r, pos_y + r)$; that is, the excursion of the regional location generates a vertex, as shown in Fig. 3.

When m = 1, the calculation of p is as shown in Eq. 4:

$$p = S/S_{pox} = 2 * (\frac{1}{4}\pi r^2 - \frac{1}{2}r^2)/\pi r^2 = \frac{1}{2} - \frac{1}{\pi} \tag{4}$$

when m = 1, the value of the minimum query accuracy is $\left(\frac{\pi/2-1}{\pi}\right) \approx 0.1817$.

4 Experimental results and analysis

4.1 Query accuracy and communication overhead

Due to the dummy locations being randomly generated, the minimum of query accuracy is uncertain. Thus, we conducted a large number of simulation experiments to

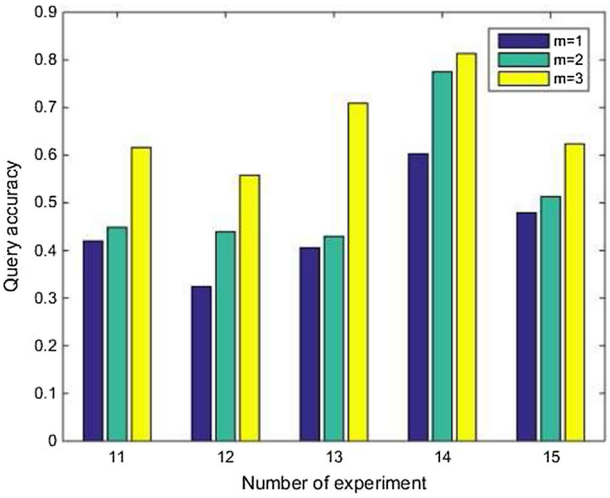


Fig. 4 Query accuracy

find the minimum approximation, and consequently verified that the influence of query radius on the accuracy of all dummy k-anonymous algorithm. Based on the location offset, the influence of the accuracy experimental set's anonymous area length is 400, while the query radius is 40. The query accuracy from 11 to 15 times under different query radii and the minimum query accuracy in these 15 experiments were calculated, as shown in Fig. 4. As can be seen from the figure, the query accuracy is random, because the selection of dummy locations is random. At the same time, it can be seen that the greater the value of m, the higher the query accuracy will be.

In a general query request, it takes 8 bytes to use latitude and longitude to represent the location information. Thus, when using n dummy locations for the query, the data submitted can be represented as:

$$req = 8n \tag{5}$$

Using the dummy location method, the data returned to the user by the server side has:

$$res = 8 \sum_{i=1}^{n} R_i \tag{6}$$

Here, R_i represents the result at the corresponding location i in the query request Q.

According to the process of generating fake addresses by the all-dummy k-anonymous algorithm, the communication overhead of the all-dummy k-anonymous algorithm is the same as that of the traditional k-anonymous algorithm when the number of locations initiated by the query is the same. The reason is that, in the process of generating dummy locations, without accounting for location information processing,

each dummy location query and normal query communication overhead is the same, which makes the number of dummy locations is the same. The two methods have the same total communication overhead.

According to the process of all dummy k anonymous algorithm to generate dummy position as we can see, the computing time of all dummy k anonymity algorithm mainly comes from the random dummy location process of generation, the generation of each dummy position requires two addition, assume that each addition operation time for a, all dummy anonymous computational time is $2(k-1)a$. It can be seen that the calculation time is related to the size of k and the single calculation time. The single calculation time is related to the server performance, and the calculation speed is relatively fast due to the addition. Therefore, the computing time of the all-dummy k anonymous algorithm can be regarded as only related to the number of dummy locations. Because the calculation is relatively simple, so the calculation time is relatively small.

4.2 Entropy of the anonymous set

The entropy of the user's location privacy information can be used as a measure of the degree of protection provided. In this article, entropy is used to compare the advantages and disadvantages of algorithms. The experimental parameter settings are as follows: the side length R of the anonymous area is 700, while the query radius r is 40. Contrast the location in the same dummy number of cases. Under the same number of dummy locations, the entropy changes of the algorithm proposed in this paper, random k-anonymity, Grid Dummy, Circle Dummy and optimal mechanism are compared; the maximum number of dummy locations is 30. Grid Dummy and Circle Dummy are respectively based on the grid and circle dummy location generation algorithms. The optimal mechanism refers to the theoretical value of the k-anonymity algorithm under the ideal state. Results are as shown in Fig. 5. As can be seen from the figure, as the dummy location entropy increases, the privacy protection effect is. The dummy anonymous algorithm achieves better performance than random k-anonymity, Grid Dummy and Circle Dummy, but worse performance than the theoretically optimal k-anonymous algorithm.

4.3 Tracking success rate

In this paper, $\boldsymbol{R_A}$ is used to represent the user's real location,$\boldsymbol{ASA}$ represents an anonymous set, and $|\boldsymbol{AS_A}|$ represents the number of elements in the anonymous set. Suppose $\boldsymbol{p}(\boldsymbol{A}, \boldsymbol{B})$ is the probability of an attacker taking the location of B in the anonymous set as the location of user A; thus, the anonymous set can be expressed as follows:

$$AS_A = \{R_B | p(A, B) \neq 0\} \tag{7}$$

The probability that the attacker can find the user's real location (tracking success rate) P_A is the probability that the size of the anonymous set $|AS_A|$ is 1; that is,

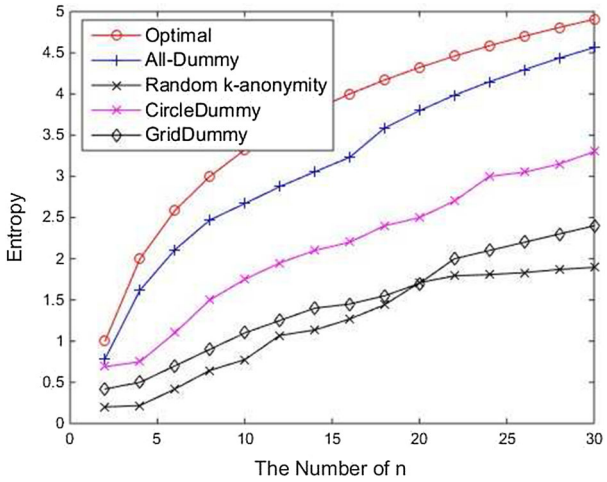


Fig. 5 The effect of the number of dummy locations on entropy

$P_A = P_r(|AS| = 1)$. When there is only one element in the anonymous set, P_A=1. The more elements are in the anonymous set, the smaller the value of A will be, and the higher the degree of user privacy protection will be.

In our experiment, we compared the all-dummy k-anonymous algorithm, the random k-anonymous algorithm, the circle-based dummy location generation method and the tracking success rate under the optimal mechanism. The experimental results are as shown in Fig. 6. From the image, it can be seen that the P_A values of the all-dummy k-anonymity algorithm are better than those of the random k-anonymity algorithm

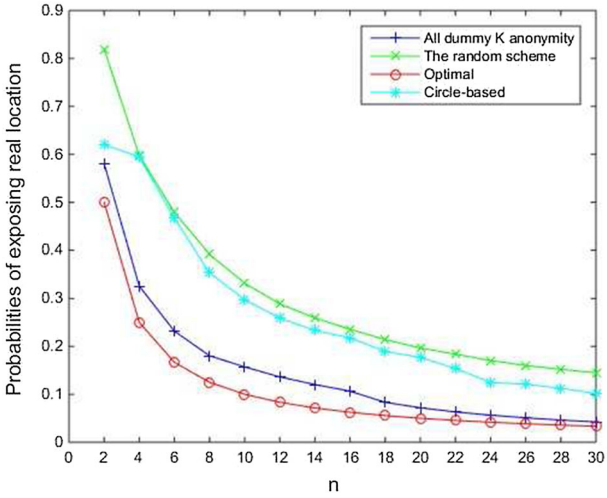


Fig. 6 Number of dummy locations vs. tracking success rate

and the circle-based dummy location generation method. Moreover, as the number of dummy locations increases, P_A grows closer to the optimal mechanism, indicating that the degree of user location privacy protection is increasing.

In this paper, the all-dummy k-anonymous location offset-based algorithm has two key advantages:

1. Even when dummy locations are generated in areas users cannot reach, the attacker cannot discern that the dummy location has nothing to do with the user's true location, because all locations are after migration; the real location used to generate a set of dummy locations may also be in an area that the user cannot reach, but the attacker cannot rule out part of the dummy location accordingly, which improves the degree of privacy protection provided for the user's location information.
2. Compared with the method of generating dummy location through caching and querying probability, the all-dummy k-anonymity algorithm based on location offset has a degree of protection of user privacy close to that of the optimal mechanism, and utilizes computing resources and communication overhead that is roughly equivalent to random k-anonymity at the same scale.

5 Conclusion

In order to improve the degree of privacy protection for user location information, this paper improves the traditional random k-anonymous algorithm, proposing an all-dummy k-anonymous algorithm based on location offset. The algorithm randomly generates k−1 dummy locations in the anonymous area; subsequently, the dummy locations are used as locating points. The locating coordinates are used for migration, with each locating point generating $\boldsymbol{m}$ new locations. In order to further improve the degree of protection provided for the user's location information, all of the generated offset locations are further away than the distance of the query radius $\boldsymbol{r}$ to query the dummy location, thereby achieving the aim of protecting users' privacy. Compared with the traditional random k-anonymous methods, this paper's proposed all-dummy k-anonymous algorithm based on location offset also generates dummy locations, but the attacker cannot determine the authenticity of the location simply because the dummy location is in a place that the user cannot reach. Moreover, because the coordinates used in the query are all dummy locations, the attacker cannot determine the user's real location.

References

1. Chen J, He K, Yuan Q (2018) Blind filtering at third parties: an efficient privacy-preserving framework for location-based services. IEEE Trans Mob Comput 17(11):2524–2535
2. Soria C, Domingo F, Josep S, David M, David J (2017) Individual differential privacy: a utility-preserving formulation of differential privacy guarantees. IEEE Trans Inform Forensics Sec 12(6):1418–1429
3. Peng T, Liu Q, Meng D (2017) Collaborative trajectory privacy preserving scheme in location-based services. Inform Sci 387:165–179
4. Niu B, Li Q, Zhu X, Cao G, Li H (2014) Achieving K-anonymity in privacy-aware location-based services. In: Proceedings of IEEE INFOCOM. Toronto, Canada: IEEE, pp 754–762.
5. Niu B, Li Q, Zhu X, Cao G, Li H (2015) Enhancing privacy through caching in location-based services. In: Proceedings of IEEE INFOCOM on computer communications. Hong Kong, China: IEEE
6. Li W, Li C, Geng Y (2020) APS: attribute-aware privacy-preserving scheme in location-based services. Inform Sci 527:460–476
7. Tan R, Tao Y, Si W (2020) Privacy-preserving semantic trajectory data publishing for mobile location-based services. Wireless Netw 26(8):5551–5560
8. Wu Z, Wang R, Li Q (2020) A location privacy-preserving system based on query range cover-up or location-based services. IEEE Trans Veh Technol 69(5):5244–5254
9. Liu J, Jiang X, Zhang S, Wang H, Dou W (2019) FADB: frequency-aware dummy-based method in long-term location privacy Protection. In: Proceedings of the 2019 IEEE 25th international conference on parallel and distributed systems (ICPADS). Tianjin, China: IEEE, pp 384–391.
10. Yang X, Gao L, Zheng J, Wei W (2020) Location privacy preservation mechanism for location-based service with incomplete location data. IEEE Access 8:95843–95854
11. Hua L, Christian SJ, Man LY (2008) PAD: privacy-area aware, dummy-based location privacy in mobile services. ACM MobiDE
12. Ni L, Tian F, Ni Q (2020) An anonymous entropy-based location privacy protection scheme in mobile social network. Eurasip J Wireless Commun Netw 2020:93
13. Liu H, Li X, Li H (2017) Spatiotemporal correlation-aware dummy-based privacy protection scheme for location-based services. In: IEEE conference on computer communications (INFOCOM). Atlanta, GA