Peer-to-Peer Networking and Applications (2019) 12:1097–1109
https://doi.org/10.1007/s12083-019-00774-8

Location protection method for mobile crowd sensing based on local differential privacy preference

Jian Wang[1] · Yanli Wang[1] · Guosheng Zhao[2] · Zhongnan Zhao[1]

Received: 29 November 2018 / Accepted: 5 June 2019 / Published online: 3 July 2019

Abstract
In view of the location privacy problem of participants in mobile crowd sensing, this paper proposes a method to protect the location of participants based on local differential privacy preference. First of all, the map is discretized and mapped from two-dimensional space to one-dimensional space by means of MHC, which can guarantee the spatial correlation, and the map is segmented based on the density of participants using genetic algorithm; Then, according to the personal privacy needs of current location, two different local differential privacy perturbation methods, RAPPOR and *k*-RR, are chosen by participants; Next, the chosen local differential privacy is used to perturb the location of each participant in the region after segmentation, and the perturbed location data are sent to the data collection server to protect the participants' locations. Finally, the simulation experiments are carried out and show that map density segmentation can reduce the privacy cost, and the method proposed in this paper is superior to the method using k-anonymous and differential privacy and the method using Hilbert and differential privacy in terms of running time and average relative error, and prove that the execution time is lower and the data availability is improved.

Keywords Mobile crowd sensing · Local differential privacy · RAPPOR · *k*-RR · Map density segmentation

1 Introduction

In recent years, large-scale and high-precision environmental sensing has become an important prerequisite for carrying out various social activities, and in the face of uncertain and large-scale sensing environment such as limited network resources, the sensing quality obtained through the pre-deployment of dense sensor nodes is difficult to satisfy actual needs. However, through ubiquitous intelligent terminals and network access, mobile crowd sensing (MCS) migrates the sensing task from the centralized platform to the distributed computing terminal across spatial and temporal dimensions [1–3], providing new ideas for large-scale and high-precision real-time sensing problems. Unfortunately, while mobile crowd sensing brings convenience to people, it also brings hidden dangers of participants' location privacy. As participants are aware of the potential possibility of personal location leakage, the number of participants is reduced. Some participants upload false sensing information or refuse to participate in sensing in order to protect their locations. This situation will reduce the quality and quantity of sensing data. Therefore, it is very important to protect the participant's location in mobile crowd sensing.

K-anonymity is currently widely used in location privacy protection, Gruteser et al. [4] introduced the concept of *k*-anonymity into location privacy, and proposed a simple and effective spatial concealment method for the first time, but participants could not set different *k* and concealment areas according to their requirements. Chow et al. [5] proposed a personalized *k*-anonymity scheme, but *k*-anonymity could not reflect the level of privacy protection accurately when the probability of user occurrence in the conceal area is not equal. Therefore, many scholars proposed the concept of location entropy, which is originated from Shannon entropy, and that is a method to measure uncertainty, which is also widely used in location privacy protection. Beresford et al. [6] proposed a privacy protection scheme based on mix-zone, but it didn't consider participants' motion mode. Palanisamy et al. [7]

✉ Jian Wang
wangjianlydia@163.com

[1] School of Computer Science and Technology, Harbin University of Science and Technology, No.52 Xuefu Road, Nangang District, Harbin 150080, China

[2] School of Computer Science and Information Engineering, Harbin Normal University, Harbin 150025, China

applied the mix-zone to the vehicle location privacy protection in the road network environment, while none of the above methods can resist the location inference attack.

Differential Privacy (DP) makes it possible to resist all kinds of attacks who have the greatest background knowledge [8, 9]. Differential privacy is a new privacy definition proposed by Dwork in 2006 for privacy leakage of statistical data-base. Andres et al. [10] proposed a location-based differential privacy extension model, and generated an anonymous location which would not be attacked within the privacy budget, however, the distance between the real location and the anonymous location could not be predicted, so the quality of LBS service could not be guaranteed. Dewri et al. [11] used the anonymous set distributed by Hilbert to obtain high-quality services, but there is no similarity between the anonymous location and the real location, so it is easy to be attacked. In addition, differential privacy and its variants to protect privacy are also used in algorithms that contain participant's location information. In literature [12], a differential privacy mechanism for location privacy protection in spatial-based group perception task was proposed. Tong et al. [13] put forward a private scheduling protocol for ridesharing services, in which participant's location information was protected under a state-of-the-art variant of differential privacy, joint differential privacy [14]. Jin et al. [15] studied the location privacy protection in the crowdsourced spectrum perception. In their work, participants' locations were protected by differential privacy, and system objectives such as task fulfillment were optimized. Nevertheless, differential privacy protection for sensitive information is always based on a premise: a trusted third-party data collector, that is, to ensure that the third-party data collector doesn't steal or leak sensitive information about participants. In practice, participant's privacy is not guaranteed even if third-party data collectors claim that they will not steal and leak sensitive information.

For these questions, a method for mobile crowd sensing based on local differential privacy preference [16, 17], namely MCS-LDPP, is proposed to protect the participant's location. While it can resist location inference attacks, it also transfers the perturbation of location data to the sensors of each participant. This can avoid sending participant's real location information to third party, and participants can individually define their privacy levels in their own sensors. Then, the location after perturbation is sent to the data collection center to better protect the location privacy of participants.

Because the location information is perturbed in the participant's sensor, reducing the privacy cost is important. First, after receiving the sensing task, participants judge which private level they are in the task's location by themselves, and then the perturbation method is selected, namely k-ary randomized response (k-RR) [18] or randomized aggregatable privacy-preserving ordinal response (RAPPOR) [19]. In the perturbation method, k-RR simplifies the process of data perturbation, which can reduce the privacy cost of the algorithm. While, if the participants have high privacy requirements, the perturbation using k-RR method makes data availability lower. Therefore, RAPPOR is selected under this condition. By using k-RR and RAPPOR, the number of participants will affect the running time. Thus, map density segmentation is conducted in this paper, and the number of participants in the corresponding region is used for perturbation according to the location of participants.

Main contributions in this paper:

1) By using local differential privacy, the participant's location is protected and perturbed in the participant's sensor, avoiding the privacy threat from the third party;
2) Participants select the perturbation method based on the personal privacy need of the current location. Considering the personal privacy need of current location, participants choose two different local differential privacy perturbation methods, RAPPOR and k-RR, so as to reduce privacy cost and meet the participant's needs;
3) When participants use k-RR or RAPPOR perturbation, the regional segmentation is executed on the server side for the participant's perturbed location. Through the regional segmentation, the privacy cost of participant's sensor is further reduced;
4) Experiments show that the method proposed in this paper can protect the participant's location, the method of map density segmentation can reduce the privacy cost and it has advantages in algorithm time and the availability of data.

2 Problem Descriptions

In mobile crowd sensing, after users participate in the sensing, participants' locations are protected by local differential privacy. The protective model of local differential privacy is fully considered the possibility that the data collector steals or leaks the participant's privacy during the process of collecting the participant's location. Participants perturb their locations firstly, and then the perturbed locations are sent to data collection server. The received location data are made statistics on the data collection server to obtain effective analysis results. That is, when the location data are statistically analyzed, the privacy information of the participants' locations can't be leaked. The definition of local differential privacy is as follows.

Definition 1 Suppose n participants, one location for each participant, given a privacy algorithm M, its domain Dom(M) and range Ran(M), if any two locations t and $t'(t, t' \in Dom(M))$ can be obtained the same output results t^* ($t* \subseteq Ran(M)$) through the algorithm M, that is, the algorithm M satisfy the following inequalities, as shown in formula (1), then M satisfies the ε−local differential privacy.

$$\Pr[M(t) = t^*] \leq e^{\varepsilon} \times \Pr[M(t') = t^*] \quad (1)$$

From definition 1, the local differential privacy ensures that the algorithm M satisfies the ε−local differential privacy by controlling the similarity of the output results of any two participants. In short, it is almost impossible to infer the true participant's location according to the privacy algorithm M. For the differential privacy, the privacy of algorithm M is defined by the neighbor dataset [20], so it requires a trusted third-party data collection server to privacy the analysis results of location data. However, for the local differential privacy, each participant can perturb their own location data independently. It means that the processing of perturbing the participant's location is transferred from the data collection server to the participant's sensor, so the trusted third party is no longer required, and the privacy attack that can be caused by untrusted third party is also exempted, as shown in Fig. 1. The participant's location is directly perturbed in the participant's sensor by local differential privacy, and then the perturbed location is sent to the data collection server. Definition 1 ensures that the algorithm satisfies ε−local differential privacy in theory, and the perturbation mechanism is required to achieve ε−local differential privacy.

3 Location Protection Algorithms for Mobile Crowd Sensing

For mobile crowd sensing, this paper proposes a method to protect participants' locations based on local differential privacy preference, MCS-LDPP. When the task publisher of mobile crowd sensing publishes the task to users, users decide whether to accept the task. Once the user accepts the task, the participant will query the server to find the current location in which region after segmentation, for getting the number of participants and the participant's id in the region after segmentation. Participants select the perturbation method according to the personal participant's privacy need of the current location. When the privacy need is low, that is, the privacy budget is high, the k-RR is used, and then the participant's location after perturbation is sent to the data collection server. On the contrary, when the privacy need is high, that is, the privacy budget is low, the RAPPOR is used, and then the participant's location after perturbation is sent to the data collection server in the same way. By perturbing the location information in the participant's sensor, the participant's location can be protected from attackers when it is in the transmission and the data collection server. Even if the attacker gets the data, it's the encrypted data, so the attacker can't get the actual participant's location, as a result, and the participant's location is protected, as shown in Fig. 2. In the map density segmentation, modified Hilbert curve (MHC) is used for the segmentation of region. The map is mapped from two-dimensional space to one-dimensional space, which can guarantee the spatial correlation. Then the map is segmented according to the density of participants by genetic algorithm. Finally, in this way, the points of region after segmentation can be obtained. Because the participants' locations are changing constantly, so regular updates are required, but this situation is not considered in this paper.

The participant's location is perturbed in the participant's sensor, thus reducing the privacy cost is important. In this paper, participants select the perturbation method according to the personal participant's privacy need of the current location, so RAPPOR or k-RR is not always used for privacy protection. Meanwhile, the privacy cost can be reduced. Because k-RR is optimal when the privacy budget is high [21], the privacy cost is reduced. In addition, when protecting the participants' locations, the number of participants is reduced through regional segmentation, further reducing the privacy cost and enhancing the practicability of this method.

3.1 Map density segmentation

The participant's location is perturbed in the participant's sensor, so reducing the privacy cost is important. In this paper, the map is divided according to the density of participants, and it can reduce the number of beacon users when RAPPOR or k-RR is used for perturbation. Consequently, the privacy cost of protecting the participants' locations can be reduced. Besides,

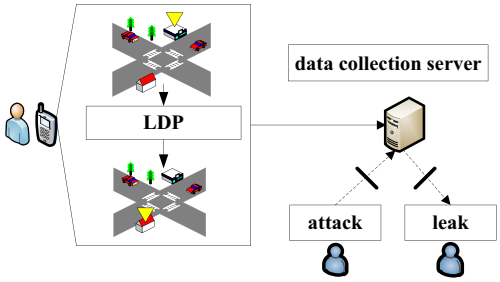


Fig. 1 Participant's location for local differential privacy protection

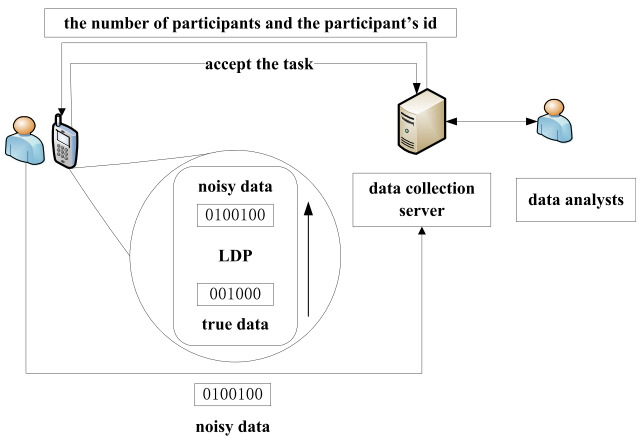


Fig. 2 Participant's location privacy protection

since the distribution of participants' locations will be sparse, after the regional segmentation, the difference of the perturbed participants' locations can be reduced, and the availability of the perturbed participants' locations can be improved. Moreover, the confusion to the attacker can be increased, and the analysis of the perturbed participants' locations can be more effective.

First, the map is divided initially to make the map discretization. The map is mapped from two-dimensional space to one-dimensional space by using MHC [22], which can guarantee the spatial correlation. Without loss of generality, it is considered that map R is a rectangular region. If and only if the number of participants in the map is greater than the predefined threshold, the map is divided into 4 sub-regions of the same size recursively, as shown in Fig. 3. It is divided by recursion according to a given threshold σ=1 until the number of participants in the region after segmentation does not exceed the given threshold σ.

Through MHC, it's easy to get a four-point tree, and there are only two possibilities for each node in the tree: leaf nodes or nodes contained four children. In order to store the tree effectively, a breadth first search tree can be built and 1 bit information is stored for each node to indicate whether the node is a leaf node, and through this way we can convert the map into serialized storage files. Figure 4 (b) shows an example of the serialized storage of a map in Fig. 4 (a). Suppose that the size of the potential participants' locations in the region is n, then the number of leaf nodes is n/σ. A quadtree with n/σ leaf nodes has $4n/3\sigma$ nodes at most. And then, the maximum amount of space required to serialize the file is $4n/3\sigma$ bits, so the storage cost of the file is O(n). Because the participant's location density in the region will not change much in a certain period of time, the MHC construction can be achieved by offline mode. In order to assign Hilbert values in one-dimensional space to each atomic unit, it is necessary to conduct a depth-first traversal of the quadtree, and Hilbert values

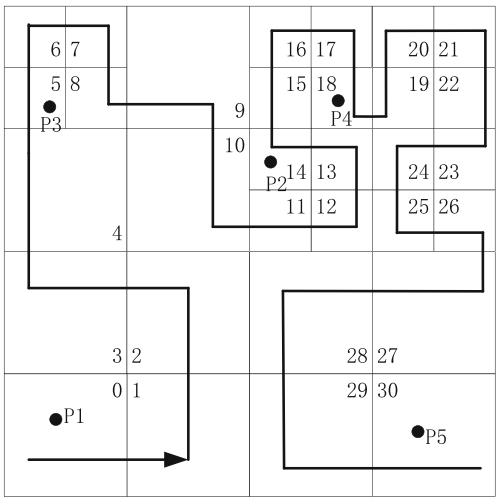


Fig. 3 MHC division

of potential participants in one-dimensional space are assigned according to the traversal order of each leaf node. As shown in Fig. 4 (a), the number below the leaf node represents the order in which each leaf node is accessed, that is, the leaf node corresponded to the Hilbert value of the atomic unit, and it is also the Hilbert value of potentially participant's location. For example, in Fig. 3, the Hilbert values of the atomic units that include the participant's location (P2, P5) are 14 and 30 respectively, and the Hilbert values of the participant's location (P2, P5) are also 14 and 30 in Fig. 4 (a). The time complexity of calculating the Hilbert value of each participant is O(n).

After the above operations, the initial segmentation of the map is obtained. The initial region segmentation sequence can be indicated as: r_1, $r_2\cdots$, r_s, and the subscript is the Hilbert value of the region. Due to the update of regional segmentation established by MHC within a certain period of time, it is necessary to traverse according to the initial regional segmentation to judge whether the potential participant is in the current region. If the potential participant is in the i-th region, then $a_i + 1$. In turn, the number of participants in each region after segmentation is received, and the sequence of participants' number in each region after segmentation can be represented as: a_1, a_2, $\cdots$, a_s.

Then, the map is segmented according to the number of participants through using genetic algorithm [23]. The map is divided into different grades on the basis of the sequence which is got by the above method, such as dense, relatively dense, medium, relatively sparse, sparse, and so on. Assuming that n regions are divided, n-1 sequence segmentation points need to be set, that is, n-1 values are selected from the sequence of subscript 1, 2, $\cdots$, s which is obtained from the sequence of participants' number in each region after segmentation. The corresponding sequence is indicated as: $D = \{d_1, d_2, \cdots, d_{n-1}\}$, where $d_1, d_2, \cdots, d_{n-1} \in \{2, 3, \cdots s-1\}$. The element values in D are arranged in ascending order, shown as: $D' = \left\{d_1', d_2', \cdots, d_{n-1}'\right\}$, where $1 < d_1' < d_2' < \cdots < d_{n-1}' < l$. If the subscripts are equal, two adjacent region cell-grids are merged. Through segmentation, the corresponding relation between the subscript set D_1, D_2, $\cdots$, D_s of s region cells and the complete sequence of subscript is as follows. It can be seen from the above that each combination of D values corresponds to a segmentation scheme, so the regional segmentation is transformed to find an appropriate subscript set D of the sequence of segmentation points, making the value of Δ minimum, as shown in formula (2):

$$\Delta = \sum_{j=1}^{s-1} \sum_{k=D_j} \left| a_k - \frac{1}{h_j} \sum_{i=D_j} a_i \right| \tag{2}$$

Where, h_j is the number of elements in set D_j.

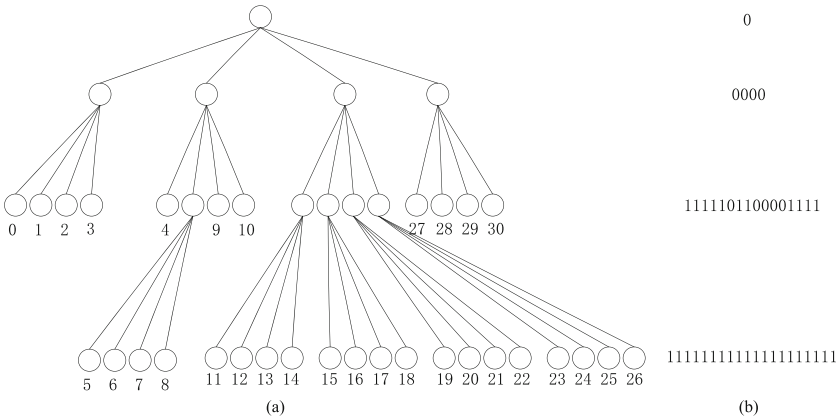


Fig. 4 MHC mapping

On the basis, genetic evolution is carried out, and individual detection, selection, crossover and mutation operations are performed. After reaching the maximum number of generations, the calculation is terminated. In the final population, the optimal solution output is selected for the individual with the maximum fitness, that is, the optimal subscript set D of the sequence of segmentation points is obtained. The algorithm for calculating the region segmentation point is described as follows.

Algorithm 1: genetic algorithm to calculate the region segmentation point.

Input: the sequence of participants' number in each region after segmentationa_1, a_2, $\cdots$, a_s, population size M, maximum generation W, control parameters of fitness calculation $\alpha$and β, adjustment parameters of selection operation γ.

Output: the region segmentation points

Step 1 G values for group M are generated randomly, where the element values in G become binary, and the encoding length of each element is $[\mathrm{lb}l]+1$, where [*] represents the integer operation, and then, the coding length of G is (s-1)($[\mathrm{lb}l]+1$). From this, the binary sequence of M group is obtained as the initial population;

Step 2 The w-th generation group is selected to determine whether the G values of M groups in the group are all within the effective range. If the subscript is not in the upper limit l, then a bit of 1 is selected randomly from the corresponding binary sequence and set to 0. If all the subscripts are in the range, take the next step;

Step 3 Follow the minimized objective expression, the corresponding $\Delta_\theta = 1, 2, \cdots, M_\theta$ of G value for M groups is calculated. Its minimum value and maximum value are $H_{\min}$ and $H_{\max}$ respectively, and $\Delta_h = H_{\max} - H_{\min}$;

Step 4 The normalized fitness is calculated, as shown in formula (3):

$$F_\theta = \left(\frac{H_{\max} - \Delta_\theta + \alpha}{\Delta_h + \alpha}\right)^\beta, \theta = 1, 2, \cdots, M \tag{3}$$

Step 5 Selection operations: verify whether $F_\theta \geq rand * \gamma$is true, if not, eliminate the θ-th individual, then verify the next individual; if the expression is true, proceed to the next step;

Step 6 The θ-th member of the population is selected to produce the next generation, and the individual itself is directly passed on to the next generation;

Step 7 Crossover operations: two crossing points are generated randomly, and two individual sequences are exchanged divided by crossing points;

Step 8 Mutation operations: A point is selected randomly from an individual, and its bit location is reversed to form the $w+1$ generation group;

Step 9 Verify if w is equal to W, if it's not equal to $w+1$, and go back to step 2; if it satisfies, go to the next step;

Step 10 In the final population, the maximum fitness individual is selected as the output of the optimal solution, in other words, the optimal subscript set G of the sequence of segmentation points is achieved.

3.2 Privacy protection for participant's location

After receiving the sensing task, participants can select the perturbation method according to the personal privacy need of the current location. In terms of different privacy budgets, the performance of the two perturbation methods, RAPPOR and k-RR, shows some difference [24]: with the privacy budget $\varepsilon = \ln k$, the RAPPOR is used for the lower privacy budget, while the k-RR is better when the privacy budget is high. Therefore, in this paper, when the privacy budget $\varepsilon < \ln k$, the RAPPOR is used, and when the privacy budget $\varepsilon \geq \ln k$, the k-RR is selected.

3.2.1 *k*-RR perturbation algorithm of participants' locations

Once the participant selects the privacy need of the current location, the *k*-RR is used when the privacy budget is high. For the *n* participants' locations in the region, the location $x_i \in \chi$, which is the *i*-th participant U_i. The value of $\chi$is obtained from the region after segmentation by algorithm 1 and $|\chi| = k$ (*k* is got from algorithm 1). When $k > 2$, we can respond immediately. For any input in the participant's location $R \in \chi$, the output of participant's location $R' \in \chi$ is shown in the following formula (4):

$$P\left(R'|R\right) = \frac{1}{k-1+e^{\varepsilon}} \begin{cases} e^{\varepsilon}, & if\ R' = R \\ 1\ , & if\ R' \neq R \end{cases} \tag{4}$$

That is, response to real results with the probability of $\frac{e^{\varepsilon}}{k-1+e^{\varepsilon}}$, and in the probability of $\frac{1}{k-1+e^{\varepsilon}}$ response to the other results, so it satisfies the ε-local differential privacy.

3.2.2 RAPPOR perturbation algorithm of participants' locations

The RAPPOR is used for perturbation when the participant's location is selected to be in the low privacy budget. Let $A = \{a_1, a_2, \ldots, a_s\}$ be whether the participant accepts a task in the region after segmentation, and *s* denotes the number of participants in the region after segmentation, which is calculated by algorithm 1 in section 3.1. For the *i*-th user, if it is the participant, a_i is set to 1; otherwise, a_i is set to 0. Let *U* be an array of s bits, and U_j represents the value of the *j*th bit in *U*. That is to say, when the *i*th user accepts the task, the bit corresponding to the user is set to 1 and the other bits are set to 0, as shown in formula (5):

$$U_j = \begin{cases} 1, & if\ j = i \\ 0, & otherwise \end{cases} \tag{5}$$

The next step is to perturb *U* which is obtained from the previous step by RAPPOR. Each bit in *U* is first perturbed by randomized response, as shown in formula (6):

$$P\left(U_j' = x\right) = \begin{cases} 0.5f, & x = 1 \\ 0.5f, & x = 0 \\ 1-f, & x = U \end{cases} \tag{6}$$

Where, $f (f \in [0, 1])$ is a system parameter that controls the privacy level. It means that values close to 1 enforce a stronger privacy guarantee. In RAPPOR, the generated U' is called the permanent random response.

Then, another perturbation is applied to each bit of U', and the instantaneous random response is obtained, denoted as S, as shown in formula (7):

$$P(S_j = 1) = \begin{cases} q, & if\ U_j' = 1 \\ p, & if\ U_j' = 0 \end{cases} \tag{7}$$

Where, $p \in [0, 1]$ and $q \in [0, 1]$ represent the probability of setting $S_j = 1$when $U_j' = 1$ and $U_j' = 0$ respectively.

At last, the instantaneous randomized response *S* is sent to the data collection server.

It can be seen from the Literature [19], when $\varepsilon = h\log\left(\frac{q^*(1-p^*)}{p^*(1-q^*)}\right), q^* = \frac{1}{2}f(p+q) + (1-f)q,$ and $p^* = \frac{1}{2}f(p+q) + (1-f)p$, the above random encoding method satisfies ε-differential privacy.

Algorithm 2: **perturbation algorithm of the participant's location.**

Input: the participant's location, and the participant's privacy need for the current location.

Output: participant's location after perturbation

Step 1 Select the perturbation method according to the personal participant's privacy need for the current location;

Step 2 If $\varepsilon \geq \ln k$, get the *k* which is the number of participants in the region after segmentation from the data collection server, and *k*-RR is used to perturb the participant's location. The perturbation steps are as section 3.2.1;

Step 3 Else get the *k* which is the number of participants in the region after segmentation and the participant's id from the data collection server, and RAPPOR is used to perturb the participant's location. The perturbation steps are as section 3.2.2;

Step 4 The participant's location after perturbation is sent to the data collection server.

4 Simulation experiments and analysis

Gowalla data set are adopted in this paper, and experimental environments are in Window 10 operating system, Intel core i5-7300 processor, and 8GB memory, and the algorithm is written in MATLAB language.

4.1 Parameter setting

In genetic algorithm, the range of population size *M* is in [200,400], and the range of maximum evolutionary algebra *W* is in [100,300]. $\alpha$is positive, generally 10^{-6}; β is a positive integer which is 1, 2, 3; $\gamma > 0$ and it is close to 1; *rand* is a random number between [0,1].

In Fig. 9 (a) and Fig. 11, *f* is from 0 to 0.4, increasing by 0.1, and $(q, p) = (0.75, 0.25)$. At this time ε is from $ln(9)$ to

ln(3.35); *f* is 0.2, and (*q*, *p*) are (0.65, 0.35) and (0.55, 0.45). In this case, ε are *ln*(2.66) and *ln*(1.38). In Fig. 7 (b) and Fig. 10, ε are set as 7, 7.25, 7.5, 7.75, 8, 8.25.

4.2 Proof of privacy

In this experiment, 750 users' locations are randomly selected, as shown in Fig. 5 (a). In Fig. 5 (a), the blue points are the selected 750 users' locations, and the horizontal and vertical coordinates are the longitude and latitude of the users' locations respectively. 40 users' locations are randomly selected from 750 users' locations as the participants' location. 40 participants' locations are distributed in different regions and the selected participants' locations are represented by red points, as shown in Fig. 5 (b).

According to the region after segmentation, the number of participants in the current region is obtained. Through the number of participants, two perturbation methods are distinguished, and the privacy budget is divided in $\varepsilon = \ln k$. The selected 40 participants' locations are protected by the two perturbation methods of local differential privacy, namely *k*-RR and RAPPOR. The participant's location after perturbation in both cases of $\varepsilon < \ln k$ and $\varepsilon > \ln k$ is verified, and it is proved that the participant's location can be protected effectively. The experimental results are shown in Fig. 6.

The 40 participants' locations after perturbation by *k*-RR are shown in Fig. 6 (a). Compared Fig. 6 (a) with Fig. 5 (b), it can be found that the number of red points increases. It is proved that the number of participants which is sent to the data collection server increases after the perturbation, so the participant's location can be protected after perturbing by *k*-RR.

Simultaneously, the 40 participants' locations after perturbation by RAPPOR are shown in Fig. 6 (b). Compared Fig. 6 (b) with Fig. 5 (b), it can be found that the number of red points increases, which is also proved that the participant's location can be protected after perturbing by RAPPOR. In addition, from the comparison between Fig. 6 (a) and Fig. 6 (b), it is observed that the red points in Fig. 6 (b) are more than those in Fig. 6 (a), which can be proved that RAPPOR is better for the participant's location protection than *k*-RR. Because in this experiment, the privacy budget of RAPPOR is lower than the privacy budget of *k*-RR, indicating that more noise is added to RAPPOR. So, the sensitive points in Fig. 6 (b) are more than those in Fig. 6 (a).

Through the above experiments, it is proved that two perturbation methods of local differential privacy can be used to protect the participants' locations.

4.3 The role of map density segmentation

Because the participant's location protection is protected in the participant's sensor, perturbation time affects electricity consumption. The longer the perturbation time is, the more electricity consumption is, as shown in formula (8). Thus, it can be seen that privacy cost is related to electricity consumption. In addition, perturbed locations are sent to the data collection server for further analysis. The greater the difference between the perturbed locations and the original locations is, the less information will be provided, as shown in formula (9). In order to obtain more accurate information, more participants may be needed, and as a result, the incentive cost of task publishers will be increased. The higher the data availability is, the lower the incentive cost will be. Therefore, the privacy cost is related to data availability, and the lower the data availability is, the higher the privacy cost will be. So, the privacy cost is as shown in formula (10).

$$e = k \cdot t \tag{8}$$

$$c = \frac{1}{l} \cdot r \tag{9}$$

$$w = \gamma \cdot e + (1-\gamma) \cdot c \tag{10}$$

Among them, γ and 1-γ represent the weight of electricity consumption and incentive cost for privacy cost. In this paper, without loss of generality, $\gamma = 0.5$ is chosen, indicating that in privacy cost, electricity consumption and incentive cost are

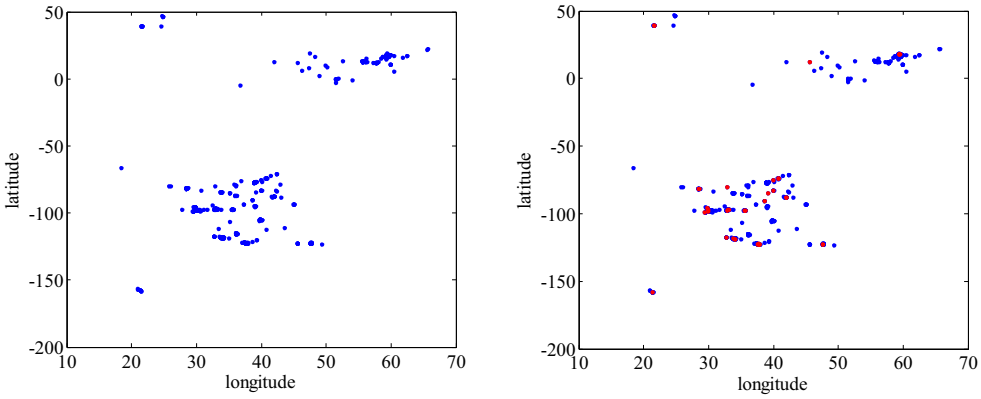


Fig. 5 Selection of initial participants location map

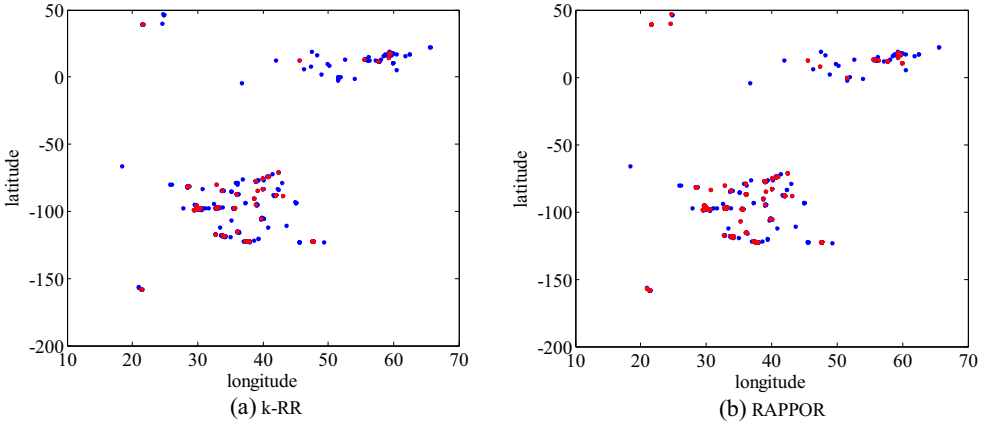


Fig. 6 Participatory user location map after disturbance

equally important. e denotes electricity consumption, t denotes perturbation time, c denotes incentive cost, and r denotes relative error, where $k = 0.1$, $l = 1$. Since the relative error range is [0, 1], there is a large difference in the value of perturbation time and relative error, and the influence of incentive cost on privacy cost is not obvious. In order to verify the role of map density segmentation, the method which does not use map density segmentation for perturbation, namely N-MDS is used for comparison to show how the map density segmentation affects the proposed scheme, and experimental results are given in Fig. 7 and Fig. 8.

From Fig. 7, with k unchanged and the privacy budget changed, it can be found that the privacy cost of N-MDS is higher than that of MCS-LDPP. Since there is no map density segmentation in N-MDS, the relative error is greater than that of MCS-LDPP, and the incentive cost c of N-MDS is greater than that of MCS-LDPP. As the number of perturbed locations is different, N-MDS is higher than MCS-LDPP in perturbation time, so the electricity consumption of N-MDS is greater than that of MCS-LDPP. In addition, it can be seen from Fig. 7 that the privacy cost w decreases with the increase of ε. This is because, with the increase of ε, the smaller the perturbation is, data availability increases and the incentive cost c decreases. Therefore, for N-MDS and MCS-LDPP, the privacy cost w decreases with the increase of ε.

From Fig. 8, with the privacy budget ε unchanged and k changed, it can be found that the privacy cost of N-MDS is higher than that of MCS-LDPP, for the same reason. In addition, it can be seen from Fig. 8 that the privacy cost w decreases with the increase of k. This is because, as k increases, the perturbation time increases and the electricity consumption e increases. As a result, for N-MDS and MCS-LDPP, the privacy cost w decreases with the increase of k.

4.4 Comparison of algorithm running time

750 users' locations are randomly selected in this experiment as the participants' locations. The algorithm proposed in this paper is compared with literature [25] and [11]. In literature [25], k-anonymity and differential privacy are used to protect the privacy of the participants' locations, represented as k + DP in the following; in literature [11], Hilbert and differential privacy is used for perturbation, finally, the perturbed location which is the minimum average distance to the original location is selected as the output, which is represented by DP below. In the experiment, in literature [25], when k-anonymous is used to protect the participants' locations, k takes the number of participants' locations in the region after segmentation; similarly, k represents the same meaning in literature [11].

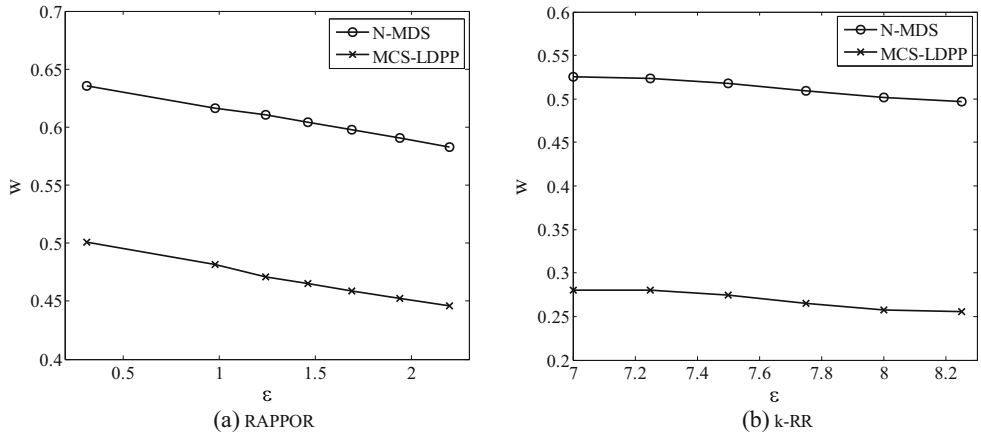


Fig. 7 privacy cost under the change of privacy budget

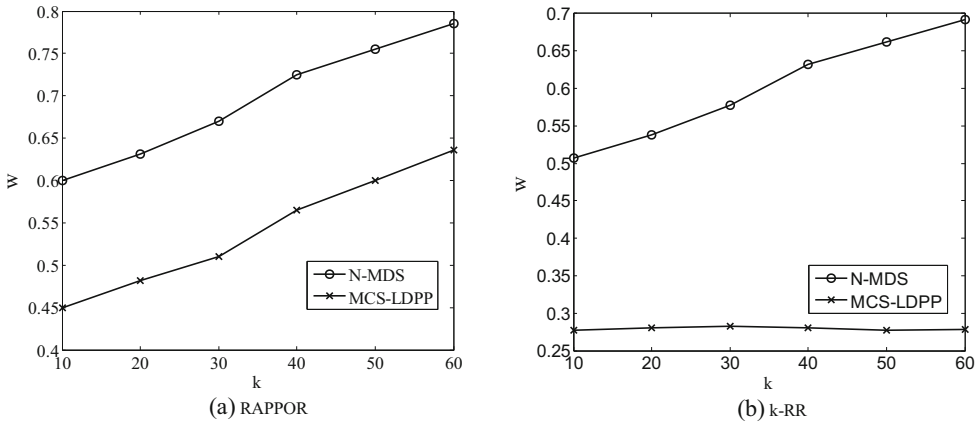


Fig. 8 privacy cost under the change of k

The experiment is divided into 4 parts. When the k is equal to 20 and the privacy budget $\varepsilon < \ln k$, the algorithm time of RAPPOR to perturb the participant's location is compared with the running time of k + DP and DP, as shown in Fig. 9 (a); when the k is equal to 20 and the privacy budget $\varepsilon > \ln k$, the algorithm time of k-RR to perturb the participant's location is compared with the running time of k + DP and DP, as shown in Fig. 9 (b); when the privacy budget is the same and the k is changed, the algorithm time of RAPPOR to perturb the participant's location is compared with the algorithm time of k + DP and DP, as shown in Fig. 10 (a); when the privacy budget is the same and the k is changed, the algorithm time of k-RR to perturb the participant's location is compared with that of k + DP and DP, as shown in Fig. 10 (b).

In Fig. 9 (a) and Fig. 9 (b), it is can be seen that when the privacy budget increases, it hasn't obvious change in the running time of MCS-LDPP which uses two different perturbation methods and the algorithm time of k + DP and DP. The reason is that the algorithm time of MCS-LDPP, k + DP and DP varies with k. And in Fig. 9 (a) and Fig. 9 (b), the algorithm running time of MCS-LDPP is lower than the running time of k + DP and DP, it means that MCS-LDPP has less privacy cost than k + DP and DP when the participant's location is protected by MCS-LDPP. The reason is that in literature [25], k-anonymity and differential privacy are used to protect the participants' locations, which are perturbed twice; in literature [11], when the value of k is a constant, the time to judge the distance between the perturbed location and the original location is invariant, at the same time, the time of perturbing locations does not vary with k. Nevertheless, the MCS-LDPP method proposed in this paper is that perturbed locations are directly output after local differential privacy, so MCS-LDPP is better in the algorithm running time.

The running time of RAPPOR is showed in Fig. 10 (a) when the privacy budget doesn't change ($\varepsilon < \ln k$) and ε = 0.98. With the increase of k, the algorithm time of MCS-LDPP, k + DP and DP increases. RAPPOR perturbs every bit of k, and the participant's location is firstly protected by k-anonymous in k + DP, and then, the participant's location which has perturbed by k-anonymous is protected by differential privacy again, and the participant's location is protected by differential privacy in DP, and then DP should consider the distance between the perturbed location and the original location, and it takes a process to choose the perturbed location which is the minimum average distance to the original location as output. So, the running time increases with the increase of k. In Fig. 10 (a), it can be found that the algorithm time of MCS-LDPP is lower than the algorithm time of k + DP and

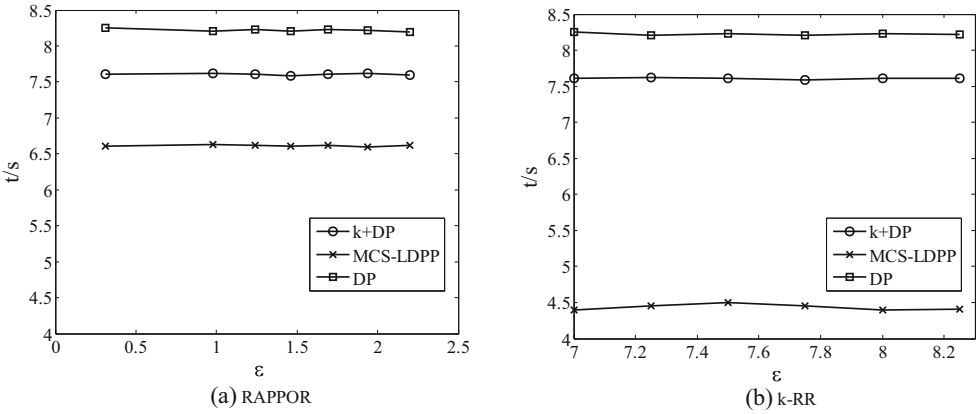


Fig. 9 Algorithm running time under privacy budget change

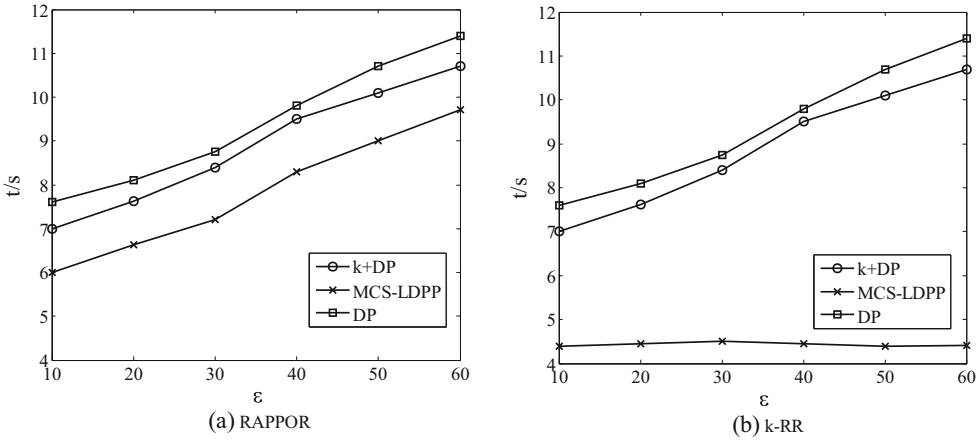


Fig. 10 Algorithm running time under the change of k

DP, and it means that MCS-LDPP has less privacy cost than k + DP and DP when the participant's location is protected by MCS-LDPP. Meanwhile, with the decline of k, the algorithm time of MCS-LDPP decreases, so it proves that when using RAPPOR to perturb the participants' locations, reducing the number of participants by regional segmentation can reduce the privacy cost.

The running time of k-RR is showed in Fig. 10 (b) when the privacy budget doesn't change ($\varepsilon < \ln k$) and $\varepsilon = 7.25$. With the increase of k, the algorithm running time of k + DP and DP increases, and the algorithm time of MCS-LDPP is basically unchanged. It is because that the algorithm time of k-RR is independent of the change of k. In Fig. 10 (b), for the same reason, the algorithm time of k-RR is obviously lower than the algorithm time of k + DP and DP, namely MCS-LDPP has less privacy cost than k + DP and DP when the participant's location is protected by MCS-LDPP. That is to say, it is proved that classification can reduce the privacy cost significantly.

From the comparison between Fig. 9 (a) and Fig. 9 (b), and the comparison between Fig. 10 (a) and Fig. 10 (b), the algorithm time of RAPPOR is lower than the algorithm time of k + DP and DP. So, it proves that when protecting the participant's

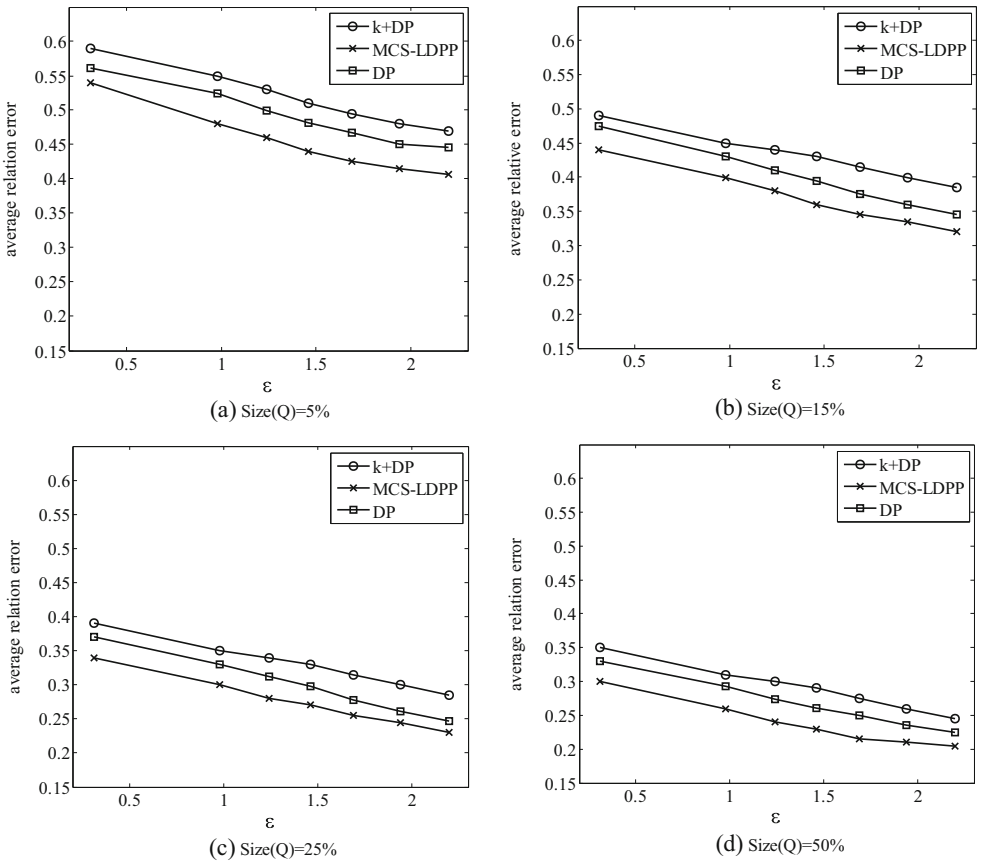


Fig. 11 Average relative error using RAPPOR

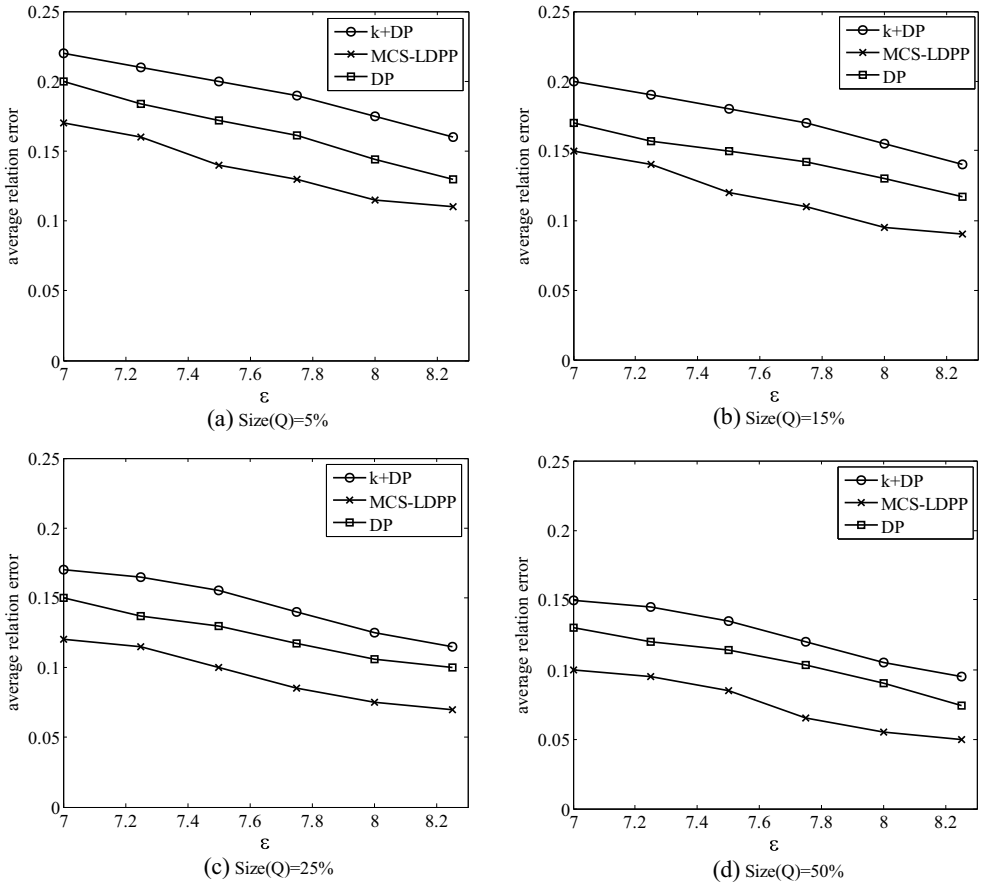


Fig. 12 Average relative error using k-RR

location, the privacy cost can be reduced by using different perturbation methods which divide through the personal participant's privacy need.

4.5 Data availability

In this paper, the relative error used in literature [26] is used to measure the availability of the participant's location, and the relative error is shown as follows:

$$\text{Error} = \frac{|Q\left(D_j^{'}\right) - Q(D_t)|}{\max(Q(D_t, s))} \tag{11}$$

The above formula calculates the error by querying the participant's location Q on the perturbed participants' locations compared to query the participant's location Q on the unperturbed participants' locations. And the parameter s is to avoid a zero in the denominator when the participant's location Q is queried.

The average relative error is tested in the case that the privacy budgets of RAPPOR and k-RR take different values respectively. The selective parameter s is set for 1% of the data set. In this experiment, the query area, shown as Size(Q) is set for 5%, 15%, 25% and 50% of the total data set, and k is equal to 750. The query is divided into different groups according to the size of Size(Q). The following figures show the experimental results. Among them, the perturbation method is RAPPOR and the results of the experiment in the RAPPOR are shown in Fig. 11, and the perturbation method is k-RR and the results are shown in Fig. 12. It's necessary to note that each of these values is the average of ten queries.

In Fig. 11, the average relative error decreases with the increase of privacy budget. The reason is that with the increase of ε, the noise added to the participant's location decreases, which makes the query more accurate. With the increase of Size(Q), under the same privacy budget, the average relative error of MCS-LDPP, k + DP and DP decreases. In addition, the average relative error of MCS-LDPP by RAPPOR is lower than that by k + DP and DP. It means that after adding noise, the data availability of participant's locations in MCS-LDPP is better than that in k + DP and DP. Because the larger Size(Q), the smaller ratio of noise data it contains, and the average relative error of DP is lower than k + DP because DP should take in account of the distance between the perturbed location and the original location, and in the same way, the perturbed location which is the minimum average distance to the original location is selected as the output in the end. And in this paper, map density segmentation reduces the distance between the perturbed location and the original location, so the average relative error of MCS-LDPP is lower than DP.

In Fig. 12, the average relative error decreases with the increase of privacy budget. Similarly, the reason is that with the increase of ε, the noise added to the participant's location decreases, which makes the query more accurate. In Fig. 12, we can find that the average relative error of MCS-LDPP by k-RR is lower than that by k + DP and DP, and it means that after adding noise, the data availability of participant's locations in MCS-LDPP is better than that in k + DP and DP. In addition, with the increase of Size(Q), under the same privacy budget, the average relative error decreases, because the larger the Size(Q), the smaller ratio of noise data it contains.

At the same time, it can be found from the comparison between Fig. 11 and Fig. 12, the average relative error of MCS-LDPP by RAPPOR is higher than that of MCS-LDPP by k-RR, because after the perturbation by RAPPOR, more noise data appears during the query.

5 Conclusions

This paper proposes a method to protect mobile crowd sensing location based on local differential privacy preference. According to the personal privacy needs of current location, participants choose two different local differential privacy perturbation methods, RAPPOR and k-RR. By using local differential privacy, the participant's location can be protected directly in the participant's sensor, avoiding the privacy threat from the third party. The map is segmented on the server, and then the number of participants and the participant's id in the participant's region are sent to the participant's sensor for participant's location perturbation. Through the segmentation of map, the perturbation privacy cost of participants' sensor is reduced. This method is proved that it has advantages in privacy protection, data availability and algorithm running time.

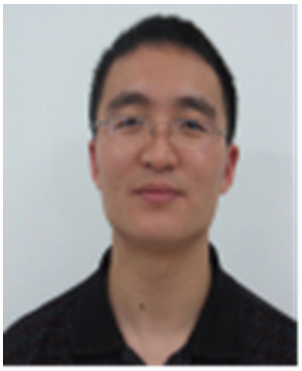
Acknowledgements This present research work was supported by the National Natural Science Foundation of China (61403109, 61202458), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20112303120007), the Heilongjiang Natural Science Foundation (F2017021), the Scientific Research Fund of Heilongjiang Provincial Education Department (12541169) and the Specialized Research Fund for Scientific and Technological Innovation Talents of Harbin (2016RAQXJ036).

References

1. Zhang X, Yang Z, Sun W et al (2017) Incentives for Mobile Crowd Sensing: A Survey[J]. IEEE Commun Surv Tutorials 18(1):54–67
2. Guo B, Wang Z, Yu Z et al (2015) Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm[J]. ACM Comput Surv 48(1):7
3. Guo B, Yu Z, Zhou X, et al. (2014) From participatory sensing to Mobile Crowd Sensing[C]. Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops, 593-598
4. Gruteser M, Grunwald D (2003) Anonymous usage of location-based services through spatial and temporal cloaking[C]. Proceedings of the 1st ACM International Conference on Mobile Systems, Applications and Services, 31-42
5. Chow CY, Mokbel MF, Aref WG (2009) Casper: query processing for location services without compromising privacy[J]. ACM Trans Database Syst 34(4):1–48
6. BERESFORD A, STAJANO F (2003) Location privacy in pervasive computing[J]. IEEE Pervasive Comput 2(1):46–55
7. Palanisamy B, Liu L (2015) Attack-resilient mix-zones over road networks: architecture and algorithms[J]. IEEE Trans Mob Comput 14(3):495–508
8. Dwork C, Kenthapadi K, McSherry F et al. (2006) Our data, ourselves: privacy via distributed noise generation[C]. Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, 486-503
9. Dwork C, McSherry F, Nissim K, et al. (2006) Calibrating noise to sensitivity in private data analysis[C]. Proceedings of the 3rd Theory of Cryptography Conference, 265-284
10. Anders M (2013) Differential privacy for location- based systems[C]. Proceedings of the 2013 ACM SIGSAC conference on Computer & Communications security, 901-914
11. Dewri R (2013) Local differential perturbations: location privacy under approximate knowledge attackers[J]. IEEE Trans Mob Comput 12(12):2360–2372
12. Jin X, Zhang R, Chen Y et al. (2016) DP Sense: Differentially private crowdsourced spectrum sensing[C]. Proceedings of the ACM Conference on Comput.Commun.Secur, 296-307
13. Tong W, Hua J, Zhong S (2017) A jointly differentially private scheduling protocol for ridesharing services [J]. IEEE Trans Inf Forensics Secur 12(10):2444–2456
14. Kearns M, Pai M.M, Roth A et al. (2014) Mechanism design in large games: Incentives and privacy[C]. Proceedings of the ACM ITCS, 403-410
15. Jin X, Zhang Y (2016) Privacy-preserving crowdsourced spectrum sensing[C]. Proceedings of the IEEE INFOCOM, 1-9
16. Kasiviswanathan S P, Lee H K, Nissim K et al. (2008) What can we learn privately[C]. Proceedings of the 49th Annual IEEE Symp.on Foundations of Computer Science, 531-540
17. Duchi J C, Jordan M I, Wainwright M J (2013) Local privacy and statistical minimax rates[C]. Proceedings of the 54th Annual IEEE Symp.on Foundations of Computer Science, 429-438
18. Kairouz P, Oh S, Viswanath P (2014) Extremal mechanisms for local differential privacy[J]. Advances in Neural Information Processing Systems, 2879-2887
19. Erlingsson Ú, Pihur V, Rappor KA (2014) Randomized aggregatable privacy-preserving ordinal response[C]. Proceedings of the 2014 ACM SIGSAC Conf.on Computer and Communications Security, 1054-1067
20. Dwork C, Lei J (2009) Differential privacy and robust statistics[C]. Proceedings of the 41st Annual ACM Symp.on Theory of Conputing, 371-380
21. Kairouz P, Oh S, Viswanath P (2016) Extremal mechanisms for local differential privacy[J]. J Mach Learning Res 17(1):492–542
22. Zhang XJ, Gui XL (2016) Jiang J H. A user-centric location privacy-preserving method with differential perturbation for location-based services[J]. J Xi'an Jiaotong Univ 50(12):79–86
23. Hu YZ, Zhang FB (2015) GanZ C, et al. QoS modeling and evaluation of mobile application service based on scene[J]. J Commun 36(Z1):110–117
24. Kairouz P, Bonawitz K, Ramage D (2016) Discrete distribution estimation under local privacy[C]. Proceedings of the 33rd Int'l Conf.on Machine Learning, 2436-2444
25. Bi XD, Liang Y, Shi HZ et al (2017) A parameterized location privacy protection method based on two-level anonymity[J]. J Shandong Univ 5(52):75–84
26. Huo Z (2018) Meng X F. A trajectory data publication method under differential privacy[J]. Chin J comput 41(2):400–412

Jian Wang, born in 1979. PhD, associate professor, Master supervisor. Member of CCF. Her main research interests include crowd sensing, cognitive network, SDN and survivability. (wangjianlydia@163.com)

Yanli Wang, born in 1994, postgraduate. Member of CCF. Her main research interests include security situation awareness and cognitive computing.



Guosheng Zhao, born in 1977. PhD, professor, Master supervisor. Senior member of CCF. His main research interests include cognitive network, and trusted computing.

Zhongnan Zhao, born in 1978. PhD, lecturer. Member of CCF. His main research interests include security situation awareness, and fault-tolerance.