An Efficient Location Privacy Protection Method for Location-Based Services based on Differential Privacy

bo Wang (B202115310017@stu.tyust.edu.cn)
Taiyuan University of Science and Technology
Hongtao Li
Shanxi Normal University
Yina Guo
Taiyuan University of Science and Technology
Xiaoyu Ren
Shanxi Normal University

Research Article

Keywords: Location-based Services, location privacy, Differential privacy, Markov model Clustering

Posted Date: April 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1504387/v1

An Efficient Location Privacy Protection Method for Location-Based Services based on Differential Privacy

Bo Wang[1], Hongtao Li[2], Yina Guo[1*], Xiaoyu Ren[2]

[1]School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China.

[2]College of Mathematics and Computer Science, Shanxi Normal University, Taiyuan 030039, China.

Author Note

*Correspondence author Yina Guo, Email: zulibest@tyust.edu.cn

Contributing authors: B202115310017@stu.tyust.edu.cn; lihongtao7758@163.com; 2658869212@qq.com;

Abstract

Location-based services (LBSs) are widely used with the rapid development of mobile devices and location technology. LBSs not only brings convenience to life, but also brings privacy threats to users. Users upload accurate location information to LBSs to obtain the corresponding services. However, uploading unprocessed location data will directly cause the leakage of users' privacy information. In this work, an efficient location protection method based on differential privacy is proposed. First of all, the users' locations information is merged based on the distance and density relationship between multiple groups of continuous locations, and a location clustering (L-cluster) algorithm is proposed to divide the continuous locations into various clusters. Then, differential privacy based location protection algorithm (DPLPA) is proposed to efficiently preserve users' location privacy, where Laplace noise conforming to the differential privacy mechanism is added to resident points and centroids in the cluster to protect the location privacy. Experimental results show that DPLPA gets high data utility and low time consumption while protecting the privacy of location information.

Keywords: Location-based Services, location privacy, Differential privacy, Markov model Clustering

1 Introduction

With the rapid development of mobile intelligent devices and location technology, various types of location-based services (LBSs) applications have brought convenience to people's life. Mobile users need to provide location information to LBSs such as the closest restaurants, hospitals or subway stations while enjoying convenient services. However, user location information is closely related to personal living habits, health, economic conditions and other private information [1]. LBSs can be used to mine and analyze the obtained users' location data, infer the users' personal preferences, and provide better-personalized services. However, directly using the unprocessed location data will cause the leakage of users' private information, and make users suffer serious location privacy problems [2].

Peer to peer technology [3] opens up a new feasible scheme to solve the problem of Internet privacy. One of the remarkable characteristics is that it is difficult for attackers to find clear attack targets, because each communication may contain many potential users. Even so, P2P network still has security issues similar to the existing Internet environment. Based on this, scholars aim by changing some strategies to enhance the privacy level. At present, most location privacy protection technologies are based on *k*-anonymity, *l*-diversity and differential privacy. *k*-anonymity and *l*-diversity protect the privacy information of users to some extent, but they cannot resist homogeneous attacks and background knowledge attacks [4]. This kind of technology is used to generalize the users' true location into a

region to realize the location protection. Dwork et al. [5] proposed a privacy protection model that can resist background knowledge attacks and provide a quantitative evaluation method, namely differential privacy. In recent years, the LBSs protection algorithms based on differential privacy have gradually become a focus of research. In this study, we proposed a continuous location privacy protection method based on differential privacy. This method protects the privacy of users' continuous location by constructing interesting areas, which not only preserves location privacy effectively, but also maximize data utility. The main contributions of this study are as follows:

(1) According to the users' access frequency to a certain location, continuous location information fusion is implemented. Moreover, according to the distance and density between locations, L-cluster algorithm is proposed to find the centroid of each cluster to replace all the locations in the cluster.

(2) The differential privacy based location protection algorithm (DPLPA) is proposed. The resident point is extracted according to whether the user's access time, access frequency and the location contain sensitive information. In addition, the privacy budget is allocated for the resident point and the cluster centroid. At the same time, the Laplace noise is added to the resident point and the cluster centroid to protect the location privacy.

(3) Theoretical analysis and experimental results show that L-cluster and DPLPA could efficiently protect location privacy for LBSs.

The rest of the paper is organized as follows. Section 2 introduces the related work of LBSs privacy protection and its major challenges; In Section 3, we give definitions of differential privacy, system structure and threat model of the algorithm; Section 4 describes the proposed L-cluster algorithm and DPLPA, and then makes a theoretical analysis of the algorithms in four parts: security, time complexity, the degree of privacy protection and data utility; Section 5 carries out simulation experiments from four aspects: clustering accuracy, the degree of privacy protection, data utility and algorithm running time; Finally, we conclude our paper and give some future perspectives.

2 Related work

Domestic and foreign scholars have done many researches on LBSs privacy protection method [6-9]. Song et al. [10] proposed a modified privacy protection scheme based on bilinear pairing theory and *k*-anonymity, in which the optimal false location was selected according to the location information. Then another author proposed a novel method of location privacy protection based on geographic semantics while satisfying k-anonymity [11], in which the candidate set was constructed by using the maximum and minimum distance multi-centers clustering algorithm, and the virtual location result set was generated according to its semantic similarity. the authors of [12] constructed a semantic and

trade-off aware location privacy protection mechanism (STA-LPPM), in which the multi-objective particle swarm optimization (MOPSO) algorithm was used to generate the optimal anonymous set to achieve the balance between privacy protection and quality of service. A blockchain-enabled framework for P2P energy trading was designed in [13], and an Anonymous Proof of Location algorithm is proposed that allows clients to choose their trading partners without revealing their real location. Zheng et al. [14] employed a dynamically adjustable *k*-anonymity (DAK) algorithm and a dynamical location privacy protection (DLPP) algorithm based on virtual locations, in which sequences were disturbed by adding and deleting some moving points. Additionally the notion of *l*-diversity and *k*-anonymity are greatly limited by data distribution and the background knowledge attacks. As a consequence, the degree of privacy protection cannot be guaranteed well.

In addition to the above methods, the LBSs privacy protection structure mainly includes location tree structure, Markov model and clustering. The main idea of location tree is to construct a tree structure according to certain rules. Prefix tree and differential privacy [15] were referred to protect trajectory data privacy, the nodes of the tree were used to store trajectory segments. Li et al. [16] considered the essential attributes associated with each location, established a hierarchical tree structure based attributes, and proposed an attribute-aware privacy-preserving scheme in location-based services. At present, a Markov model was mainly given to simulate the temporal correlation between users' real locations and predict the next possible location according to the transition probability of each location. Yuan et al. [17] proposed a new location privacy protection method for cloud of things system, in which Markov model was used to analyze the users' mobile behavior. The proposed location hiding algorithm met the users' privacy requirements by expanding the size of the area. Partovi et al. [18] modeled a Markov decision process and introduced a new location privacy measurement method to ensure that the users' specified privacy level could be achieved in an infinite time range. The *k*-anonymity method mainly used in Wu et al. [19] to enhance privacy protection, and used clustering technology to group users by learning their trajectory data. A graph based trajectory data representation model [20] was proposed, the similarity between trajectories was calculated using the measurement method based on edges and vertices, and similar trajectories were clustered and identified based on paths. Clustering can represent the users' activity rules in some time, and can remove the location with low access frequency, so it has high flexibility.

To resist the inference attacks, differential privacy become a mainstream method due to good privacy protection performance. By adding random noise to the original query results, adding or deleting a part of data in the datasets will not affect the query results. Therefore, it is difficult for attackers to infer the real data through multiple queries to

achieve privacy protection. Chen et al. [21] applied the differential privacy protection method to location data protection for the first time. By adding random noise to the location data, the users' location was confused. Hu et al. [22] considered personalized security requirements of different users (LPPA-PSRD), which realized location protection based on the users' historical GPS (global positioning system) trajectory data and the natural attributes of the location. However, it had massive computation, and the accuracy of user sensitivity evaluation needs to be improved. Thus, Wang et al. [23] proposed a privacy-protected social tie mining (P-STM) method, which could find their social connections from users' daily trajectories, and offered an indicative dense region (IDR) to calibrate personal daily trajectories. At the same time, a clustering analysis method for spatiotemporal sequence data [24] was proposed, which provided the basis for privacy protection by constructing continuous time regions, and the data publishing mechanism was also offered to resist inferential attacks, but it mainly distributed the offline group location data, and could not update other relevant information. In [25] a new framework PrivSem was represented, which combined *k*-anonymity, *l*-semantic diversity and different privacy. Compared to [26] which constructed a location search tree based on the relationship between records and added Laplace noise, and then used an exponential mechanism to select *k* frequently accessed records. Both of them provided privacy guarantees on the location privacy. However, setting the non-sensitive location as the sensitive location will increase the cost of privacy protection.

3 Preliminaries

3.1 Definitions

Definition 1 (Adjacent datasets)**.** Suppose that the datasets have the same attribute structure, and there is only one record difference between them, that is $|D \Delta D'|=1$, datasets D and D' are called adjacent datasets.

Definition 2 (Differential privacy). There is a random algorithm A and all possible outputs of A are P_A. For any two neighboring datasets D and D', and any subset S_A of P_A, if algorithm A satisfies the following conditions:

$$\Pr[A(D) \in S_A] \le e^{\varepsilon} \Pr[A(D') \in S_A]$$

$$\forall t \in Range(A), D;\ D': \frac{\Pr[A(D)]}{\Pr[A(D')]} \le e^{\varepsilon} \tag{1}$$

A satisfies ε - differential privacy

Then algorithm A provides ε-differential privacy protection, where parameter ε is called privacy protection budget. The larger the ε is, the higher the data availability is, and the lower the degree of privacy protection is; on the contrary, the lower the data availability is, and the higher the degree of privacy protection is.

Definition 3 (Privacy budget). Let d is a positive integer, D is a set of data sets, $f: D \to R^d$ is a function. The function sensitivity represented by Δf has the following definition: $\Delta f = \max \left\| f(D) - f(D') \right\|_1$, where $\|.\|_1$ is the Manhattan distance.

Definition 4 (Laplace mechanism). Given dataset D, there is a function $f: D \to R^d$, the sensitivity is Δf, then the random algorithm $M(D) = f(D) + Y$ provides ε-differential privacy protection. Where $Y \sim Lap(\Delta f / \varepsilon)$ is the random noise and obeys the Laplace distribution with the scale parameter $\Delta f / \varepsilon$. The function is in Eq.(2):

$$\boldsymbol{A}_f = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \ satisfies \ \ \varepsilon - DP \tag{2}$$

Laplace mechanism can protect differential privacy by adding Laplace noise to query results. Note that the location parameter is 0 and the scale parameter is $P(b)$ of b, then the probability density function is Eq.(3):

$$\Pr(\mu) = \frac{1}{2b} e^{-\frac{|\mu|}{b}}, \mu = (r, \theta), \theta \in [0, 2\pi] \tag{3}$$

where r is the distance of m_0 from m_1, and θ is the angle that the line m_0 and m_1 forms with respect to the horizontal axis of the Cartesian system.

Definition 5 (Interesting areas). Set the distance threshold to E. Continuous location $T_1=\{m_1, m_2, ..., m_n\}$, if $dis(m_n, m_{n+1}) \leq E$, the region formed by the sequence of moving continuous locations from location m_t to m_{t+1} is the user's interesting areas, where E is the maximum distance threshold required to form the interesting areas, and $d(m_n, m_{n+1})$ is the distance between two locations.

Definition 6 (Data utility). Data utility is measured as shown in Eq.(4):

$$U = \sqrt{\frac{\sum_{i \in R} \rho_i - \rho_{i'}}{|R|}} \tag{4}$$

Where R is the number of clusters, ρ represents the density of each cluster.

3.2 LBS system model

The LBS system structure of this study is shown in Fig.1, mainly including client, privacy protection processor, the third trust-less severs and location service provider. The client mainly obtains the users' location data through GPS, and stores the location data in the continuous location database. Privacy protection processor is divided into clustering module and continuous location protection module. The clustering module divides the users' location according to distance and density. The continuous location protection module provides differential privacy protection, and stores

the protected data on the third severs. The third trust-less severs is peer-to-peer severs. The transmission of location information is scattered among nodes without going through the centralized server, which greatly decreases the likelihood of eavesdropping and leakage of users' privacy information. Location service providers mainly query from the database and get feedback information.

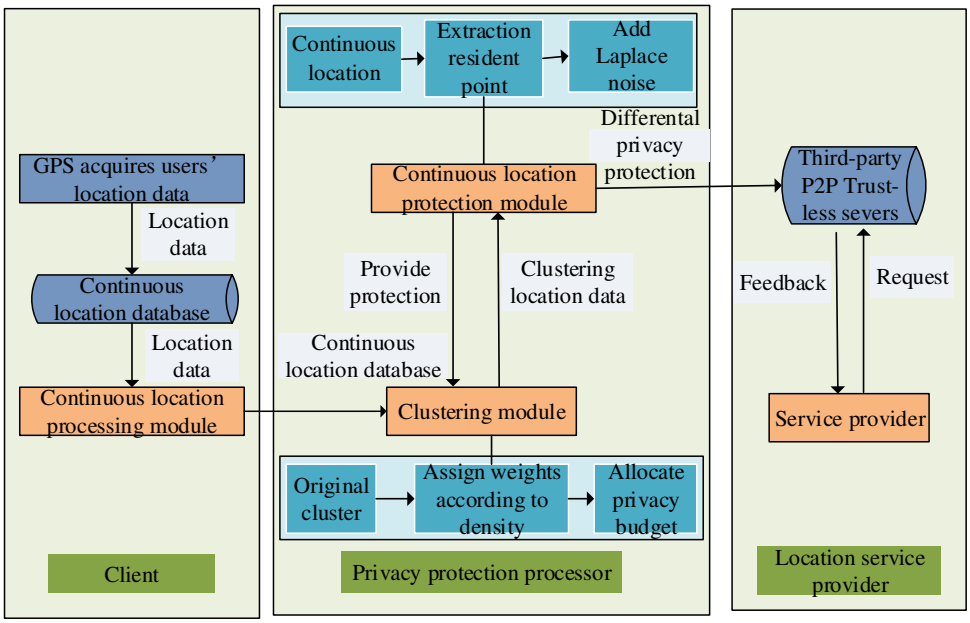


Fig. 1 LBS system structure

The system mainly adopts the fully distributed architecture and peer-to-peer network communication mode. All participants can provide the function of relay forwarding, so as to hide the communication participants in many network entities, greatly improve the flexibility and reliability of anonymous communication, and provide better privacy protection for users. It also has better robustness and invulnerability than the traditional client/server (C/S) network [27]. However, there are still some problems of privacy disclosure when users releases location information, to address the problem of users' location privacy leakage, a continuous location protection method based on differential privacy is proposed in this paper. Firstly, the users' location data is obtained by GPS, and the users' continuous location is simplified according to the location access frequency. Secondly, based on the distance and density between locations, the L-cluster method is used to cluster, the location is divided, and the clustering centroid is obtained. Lastly, the resident points are extracted by DPLPA, and Laplace noise is added to the resident points and centroids. The protected data is stored in the database for query by the location service provider.

3.3 Threat model

It is supposed that attackers will attack the users' location data from time to time. Many location-based service providers provide different security guarantees. Once these location-based service providers are attacked, users' location data and other personal information will be leaked. Based on this assumption, the threat model is proposed, as shown in Fig. 2. Smartphones, computers and other personal devices obtain users' accurate location data through GPS, and upload these data to the continuous location database. Then, the data in the continuous location database is protected by the privacy protection processor. It mainly includes three modules: continuous location processing module, clustering module and continuous location protection module. Submit the protected data to the location service provider to obtain services. From this process, the users' location data is stored in three parts: location database, privacy protection processor and location provider. Attackers can attack these three parts to obtain users' privacy information.

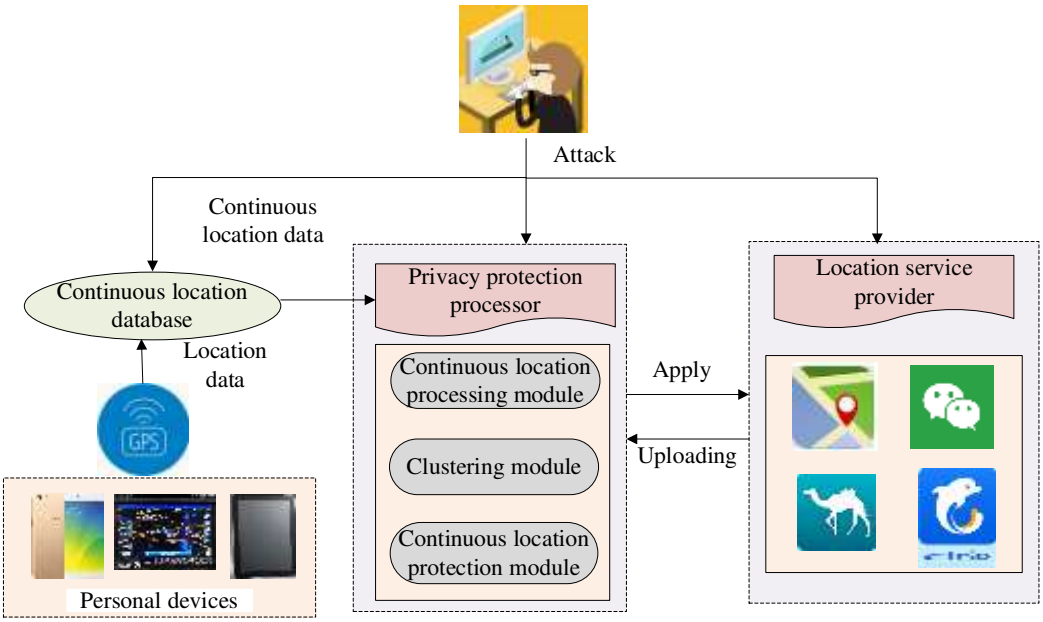


Fig. 2 Threat model

4 Differential privacy-based location protection method for LBS

4.1 User interest region construction based on clustering

In order to better protect the users' continuous location data, firstly, the users' continuous location is simplified. Summarize the users' continuous mobile locations over a period of time, and each group of records represents the users' mobile location data of the day, as shown in Table 1. Next, according to the continuity of the location in time, the users' moving continuous location is formed, and one of the locations can appear multiple times in multiple groups

of continuous locations. As shown in Fig.3, the solid dot represents the location of the user, and the line between the two locations represents the users' moving route. Finally, the users' access times to an accurate location in continuous locations are counted, and the locations with access times less than threshold h are removed, as shown in Table 2. In this way, the users' continuous location can be reduced. The result is shown in Fig.4.

For the reduced location data, it is necessary to construct interesting areas. The main method is as follows. In the continuous location, the location whose distance between locations is less than *E* is divided into the same interesting areas, and the result is shown in the dotted circle in Fig.5(a). The centroid of each interesting areas is obtained and represented by a red five-pointed star (as shown in Fig.5(b)). The centroid replaces other locations in the interesting area to form new continuous location data (as shown in Fig.5(c)).

Table 1. Original continuous location database

ID	Continuous position
1	*a→b→d→f→h→j→l*
2	*a→f→i→j→r→s→w*
3	*a→i→j→r→s→w*
4	*a→i→o→v→w*
5	*a→g→o→r→w*
6	*a→g→i→o→v→r→w*

Table 2. Location access statistics

ID	Location	Frequency	ID	Location	Frequency
1	*a*	*6*	2	*f*	*6*
3	*g*	*2*	4	*h*	*1*
5	*i*	*4*	6	*j*	*3*
7	*o*	*3*	8	*q*	*1*
9	*r*	*4*	10	*s*	*2*
11	*v*	*2*	12	*w*	*6*
13	*s*	*1*			

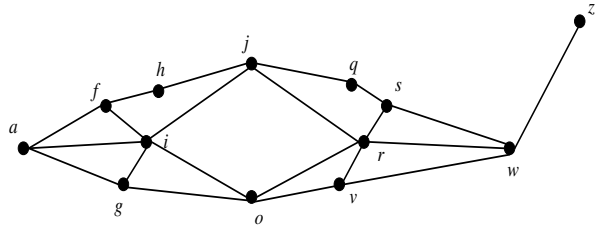


Fig. 3 Users' motion mode

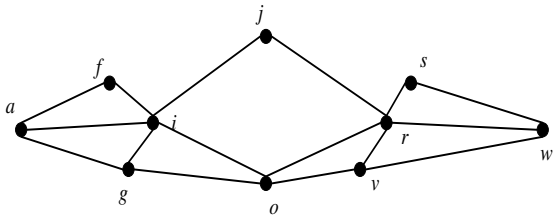
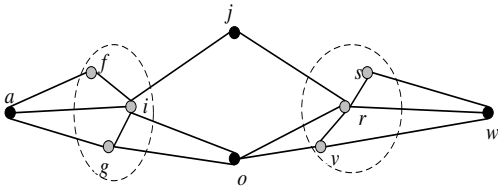
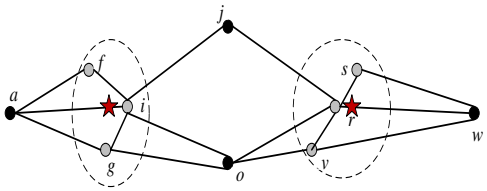


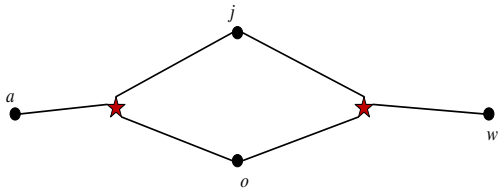
Fig. 4 Motion mode after location reduction



(a) Fusion of regions of interest



(b) Find cluster centroid



(c) Users' clustering centroid instead of region of interest

Fig. 5 Fusion continuous location formation process

Combined with the users' interesting areas, the continuous location fusion algorithm L-cluster is adopted in this study, which in order to solve the problem of location redundancy and time-consuming of density based spatial clustering of applications with noise (DBSCAN) [28]. The basic idea is as follows. Calculating the distance between the location and the adjacent location for each location. If the distance between the two locations is less than E, the two locations are divided into the same cluster. Otherwise, they belong to different clusters. This algorithm mines the users' activity range within a certain distance, and uses the centroid c of cluster to represent this area. At the same time, other location points in this area are deleted from the continuous location to avoid location redundancy. Here, the centroid is used to replace other locations to further protect users' privacy.

Algorithm 1: L-cluster algorithm

Input: Original continuous locations $T=\{m_1, m_2,\ldots, m_n\}$, maximum distance E

Output: Compact continuous location T'

1. **while** $m \in T$ do //for location t in continuous location

2. **if** m clustered is false then

3. Locations(m) = RangeQuery(m, d). //the distance between the query location m and its surrounding location

5. **else if** $d(m_n, m_{n+1}) < E$ then //if the distance between the two locations is less than the maximum distance E

6. merge m with cluster C. //location m is divided into cluster C

7. **else** m = location //otherwise, m is an independent location

8. **else if** m clustered is true //m has formed a cluster

9. **if** $d(m_n, m_{n+1}) \geq d(m_n, m_{n-1})$ //if the distance between the current location and the next location is greater than or equal to the distance between the current location and the nearest location in other clusters

10. return m. //then location m belongs to the current cluster

11. **else**

12. $m_n \in m_{n-1}$ cluster. //location n belongs to the cluster of the previous location

13. ρ is the density of clustering. //calculate the density of clusters

14. $\omega_m \leftarrow$ private weight allocation(ρ);

15. ε_m = Privacy Budget Distribution(ω_m);

16. **for** each location in T do

17. create a new cluster. //repeat the above operation to create a new cluster

18. **end for**

19. **for** each cluster C do

20. calculate the centroid c.

20. replace other locations in the cluster C with centroid c.

21. **end for**

22. return T' . //form a new compact continuous location T'

L-cluster algorithm is used for clustering and dividing the densely distributed regions, which includes four modules. The first module is from Step 3 to Step 7. It mainly judges whether the current location m belongs to a cluster, queries the distance between the current location and other locations, then compares the distance between them and the distance threshold E. If it is less than E, cluster C will be formed. Otherwise, m is an independent location; The second module judges the distance between m_n and m_{n-1} and the distance between m_n and m_{n+1} for the clustered location. If the distance between m_n and m_{n+1} is long, m_n still belongs to the current cluster, otherwise it belongs to another cluster, mainly reflected in Step8-12; The function of the third module is to allocate the privacy budget for each cluster according to the density, as shown in Step 13-15; The function of the fourth module is to calculate the centroid c of each cluster, and use the centroid to replace other locations in the cluster to form a new continuous location as the publishing location (step 19-20).

4.2 Location privacy protection algorithm based on differential privacy

Differential privacy continuous location protection algorithm needs to extract the long-time residence points, high-frequency access location points containing users' sensitive information, and define them as the resident points. For the long-time resident points, the resident time between two locations is primary considered. If it fits $t(m_i) - t(m_{i-1}) \geq t_{\text{time}}$, it is defined as the time resident points. For the high-frequency access points, the access frequency to the location

is considered. If it fits $f(m_j) - f(m_{j-1}) \geq t_{fre}$, it is defined as the frequency resident points. For the location that contains user' sensitive information, it is defined as the resident points of sensitive location. Finally, Laplace noise is added to the resident points. According to this idea, DPLPA is proposed.

Algorithm2: Differential Privacy based Location Protection Algorithm (DPLPA)

Input: Compact continuous location T', time threshold t_{time}, frequency threshold t_{fre}, distance threshold t_{dist}, sensitive attributes t_{sen}, centroid c

Output: Protection continuous location T''

1. $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$; //allocate privacy budget
2. c'= Noisy $P(\varepsilon_1)(c)$. //adding Laplace noise to centroid c
3. $i = 1$.
4. **if** $d(m_i, m_{i-1}) < t_{dist}$ and $t(m_i)$-$t(m_{i-1}) \geq t_{time}$ then
5. $m_i \in$ {long-stayed points}. //the location belongs to the time resident point
6. **end if**
7. $m_i' = \text{Noisy}P(\varepsilon_2)(m_i)$. //add Laplace noise to this location
8. $j = 1$.
9. **if** $d(m_j, m_{j-1}) < t_{dist}$ and $f(m_j)$-$f(m_{j-1}) \geq t_{fre}$ then
10. $m_i \in$ {frequency points}. //this location belongs to the frequency resident points
11. **end if**
12. $m_j' = \text{Noisy}P(\varepsilon_3)(m_j)$. //add Laplace noise to this location
13. $k = 1$.
14. **if** $d(m_k, m_{k-1}) < t_{dist}$ and m_k include t_{sen} then
15. $m_k \in$ {sensitive points}. //this location is a sensitive resident point
16. **end if**
17. $m_k' = \text{Noisy}P(\varepsilon_4)(m_k)$.//add Laplace noise to this location
18. return (m_i', m_j', m_k', c').
19. return T''.

The main content of DPTRPA is to extract the resident points and add Laplace noise which is consistent with the

differential privacy mechanism to the users' simplified continuous location T', which mainly includes four parts. The first part is Step 2, the primary function is to add Laplace noise to the centroid; The second part is Step 3-7, the primary function is to extract the time resident points and add Laplace noise to it; The third part is Step 8-12. The main function is to extract the frequency resident points and add Laplace noise to it; The fourth part is from Step 13-17. The main function is to extract sensitive resident points and add Laplace noise to them.

4.3 Algorithm analysis

4.3.1 Security analysis

Querying the location information may lead to the problem of location data leakage. To avoid this problem, a differential privacy protection method is adopted. There are two main points. (1) Differential privacy assumes that attackers have the full background knowledge, that is, the attacker knows all information except the original data. The differential privacy method can resist the maximum background knowledge attacks; (2) Differential privacy defines the degree of privacy protection through a strict mathematical model, which ensures that the ratio of query results on adjacent datasets is less than or equal to e^{ε}.

Differential privacy can protect the data by adding random noise to the input data, then output results. In this paper, Laplace noise is added to the users' location data to protect it. In DPLPA, noise that obeys Laplace distribution $P(b)$ distribution is added to the clustered data to make the noise result satisfy the differential privacy constraint. That is, the differential privacy protection is satisfied. The proof process is as follows.

It is proved that the probability density function $\boldsymbol{Pr}(\mu)=\frac{1}{2b}e^{-\frac{|\mu|}{b}}$ of Laplacian mechanism is known. x and y represent two different positions, probability density function about Pr_x is $A_m(\mathrm{x}, f, \varepsilon)$, the probability density function about Pr_y is $A_m(\mathrm{y}, f, \varepsilon)$, for a certain output value Z, there is:

$$\begin{aligned}\frac{\mathrm{Pr}_x(Z)}{\mathrm{Pr}_y(Z)} &= \prod_{i=1}^{k}\frac{e^{-\frac{\varepsilon|f(x)i-Zi|}{\Delta f}}}{e^{-\frac{\varepsilon|f(y)i-Zi|}{\Delta f}}} \\ &= \prod_{i=1}^{k} e^{\frac{\varepsilon(|f(y)i-Zi|-|f(x)i-Zi|)}{\Delta f}} \\ &\le \prod_{i=1}^{k} e^{\frac{\varepsilon(|f(x)i|-|f(y)i|)}{\Delta f}} \\ &= e^{\frac{\varepsilon\|f(x)-f(y)\|_1}{\Delta f}} \\ &\le e^{\varepsilon}\end{aligned} \tag{4}$$

Where $\|\cdot\|_1$ represents the first-order normal form distance. According to the definition of differential privacy, DPLPA satisfies the ε-differential privacy.

4.3.2 Time complexity analysis

It is assumed that there are $I]$ records in location data. The algorithm in this study involves two modules, clustering module and continuous location protection module. Therefore, the time complexity of the L-cluster algorithm and DPLPA is analyzed.

L-cluster algorithm is mainly used to cluster the regions with dense location distribution, which includes four parts. First, the location is divided into clusters according to the distance between the current location and its previous location, the current location and the following location. Its time complexity is $O(n)$; Second, according to the distance, the method can judge whether the clustering locations need to be reclassified, and its time complexity is $O(2n)$; Third, calculating the weight of each cluster according to the density and allocate the privacy budget. The time complexity is $O(n)$; Fourth, calculate the centroid c of each cluster, and use the centroid to replace other locations in the cluster. The time complexity is $O(n)$.

The main content of DPLPA is to extract different types of resident points and add Laplace noise to the continuous location T', which mainly includes four parts. First, to add Laplace noise to the centroid, and the time complexity is $O(n)$; Second, the time resident points are extracted according to the time of the access location, and Laplace noise is added to it, and the time complexity is $O(n)$; Third, the frequency resident points are extracted according to the access frequency of the location, and Laplace noise is added to it, and the time complexity is $O(n)$; Fourth, extract the sensitive resident points according to whether the users' sensitive information is included in the location, and add Laplace noise to it, and the time complexity is $O(n)$.

Generally speaking, the time complexity of this work is: $O(n)+O(2n)+O(n)+O(n)+O(n)+O(n)+O(n)+O(n)\approx O(n)$.

4.3.3 Degree of privacy protection

According to Laplace probability density function, with the increase ε, the smaller the noise is, the smaller the privacy protection is; with the decrease ε, the larger the size is, the greater the degree of privacy protection is. In DPLPA, different locations use different ε for continuous location privacy protection, and provide different degrees of privacy protection.

4.3.4 Data utility analysis

It refers to the data utility after adding Laplace noise and the gap between the processed data and the real data. Data utility is analyzed through Eq.(4) in Definition6. According to it, there are two main factors to affect data utility. The first is the number of clusters $|R|$. $|R|$ is inversely proportional to U. That is, the more clusters are, the smaller the value of U is, and the higher the data utility is. Because with the increase of the number of clusters, the higher the similarity between the simplified continuous location after clustering and true locations of the user is, the higher the authenticity of the data is. The second is clustering density ρ. Clustering density represents the number of locations in the same cluster. To some extent, it can replace the distance between locations in clustering. ρ is proportional to U. Because with the increase of ρ, there are more real users' locations that can be replaced by the cluster centroid, which makes the difference between the simplified results of the distance class and the real data larger.

Because of the continuous location reduction before the location clustering, the locations with lower users' access frequency are reduced. That is, the clustering density is minimized. Therefore, DPLPA can reduce information loss and improve data utility.

5 Experimental results and analysis Response Rates

5.1 Experimental Setting

In our experiments, DPLPA and L-cluster algorithms are implemented in Python and runs on Windows10 platform with 3.6GHz CPU and 8.00 random access memory (RAM). The datasets used in experiment are real datasets of Geolife [29] and Gowalla [30]. Then we compared DPLPA with P-STM algorithm and LPPA-PSRDU algorithm. The performance of proposed algorithm is judged from four aspects: the accuracy of clustering, the degree of privacy protection, data utility and algorithm running time.

5.2 Analysis of experimental results

5.3 Accuracy of clustering

The accuracy of clustering method is evaluated by comparing the recall, precision and F-measurement of L-cluster algorithm, K-means [31] algorithm and DBSCAN algorithm.

As shown in Fig.6, L-cluster algorithm is superior to K-means algorithm and DBSCAN algorithm in recall rate, precision rate and F-measurement value.

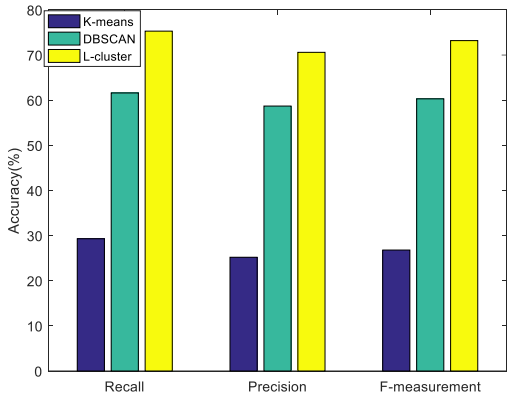
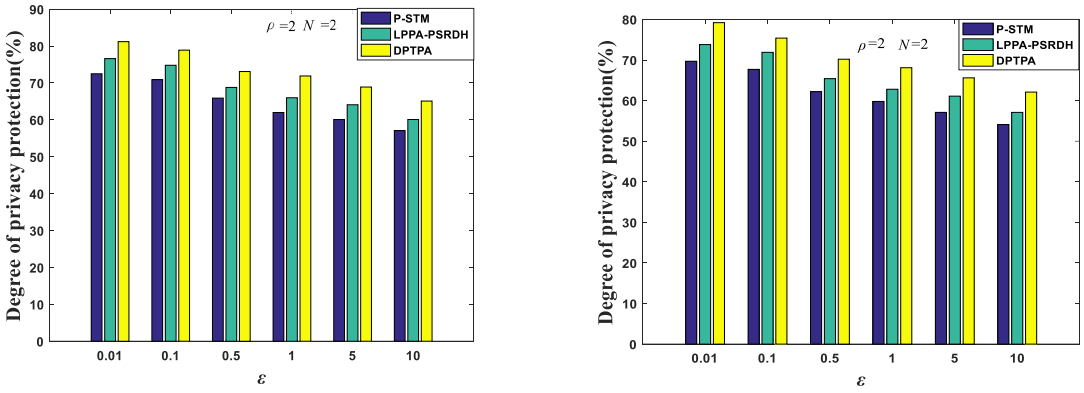


Fig. 6 Clustering accuracy

5.4 Degree of privacy protection

We analyzed the effect of privacy budget ε, cluster density ρ and the number of locations to be protected N on the degree of privacy protection. The effect of privacy budget ε on the degree of privacy protection is analyzed as shown in Fig.7.

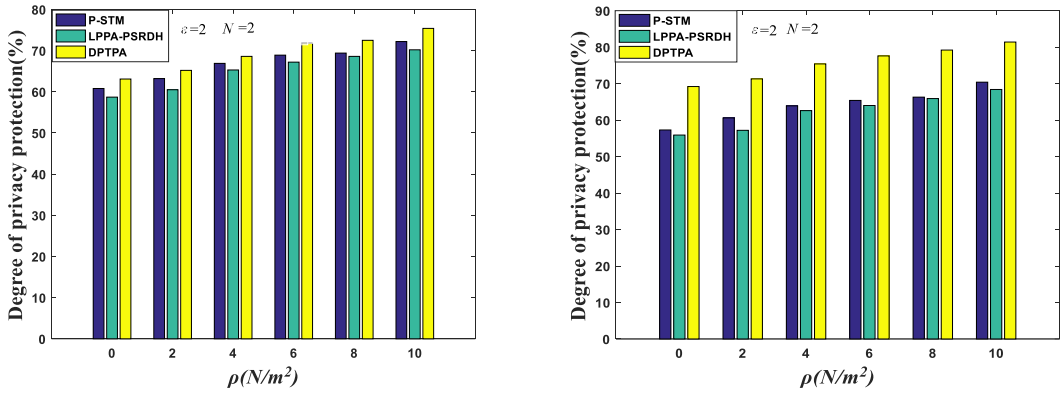


(a) Geolife dataset (b) Gowalla dataset

Fig. 7 Impact of privacy budget ε on privacy protection

The degree of privacy protection decreases with the increase of ε, which is inferred from Laplace probability density function. DPLPA has the best degree of privacy protection, followed by P-STM algorithm, and LPPA-PSRDU algorithm is the worst.

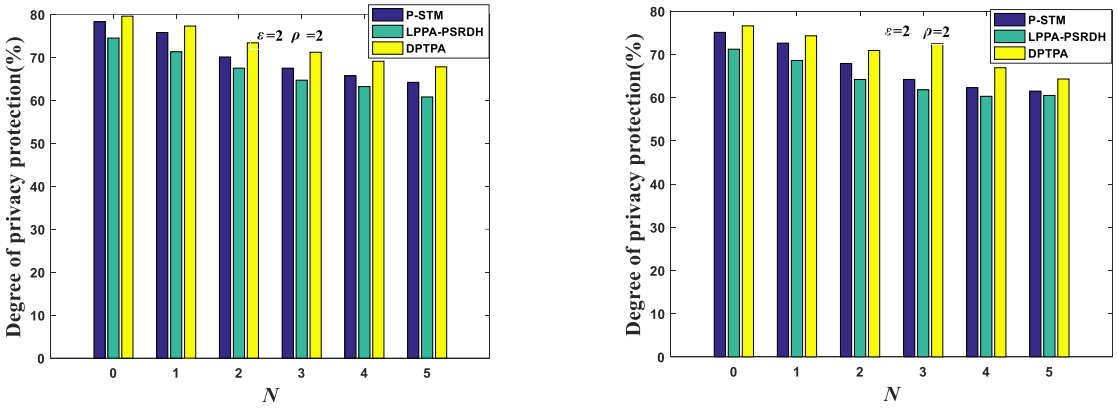
The impact of clustering density ρ on the degree of privacy protection is analyzed as shown in Fig.8. The degree of privacy protection increases with the increase of clustering density ρ. There is one centroid generated and all the locations are replaced by a unique centroid, which makes the degree of protection better. DPLPA has the best privacy protection effect, followed by LPPA-PSRDU algorithm, and P-STM algorithm is the worst.



(a) Geolife dataset (b) Gowalla dataset

Fig. 8 Impact of Clustering density ρ on privacy protection

The impact of the number of locations to be protected N on the degree of privacy protection is analyzed as shown in Fig. 9. The degree of privacy protection increases with the decrease of the number of locations to be protected N. The larger the size is, the worse the degree of privacy protection is. Similarly, the degree of privacy protection of DPLPA is the best.

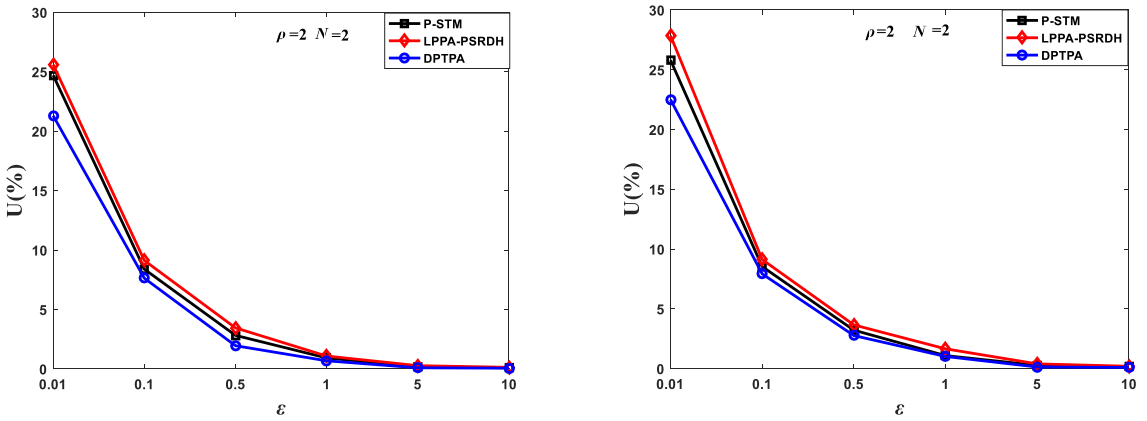


(a) Geolife dataset (b) Gowalla dataset

Fig. 9 Impact of the number of locations to be protected N on privacy protection

5.5 Data utility

By comparing DPLPA with P-STM algorithm and LPPA-PSRDU algorithm, the advantages of DPLPA in data utility are reflected. P-STM algorithm considered the similarity of the trajectory and calibrated the trajectory by using it. LPPA-PSRDU algorithm calculated the rules of users according to the historical GPS track data of users, and added noise to them by combining natural and social attributes.



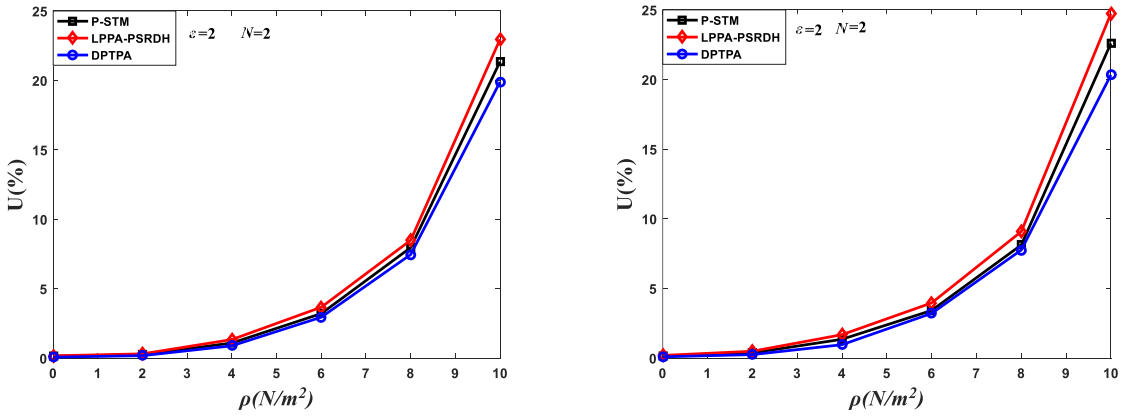
(a) Geolife dataset (b) Gowalla dataset

Fig. 10 Impact of privacy budget ε about three algorithms on U

The impact of privacy budget ε on U is analyzed as shown in Fig. 10. U increases with the decrease of ε. The data utility of LPPA-PSRDU algorithm is the worst, because it considers many factors that affect the location, and loses part of the data integrity. Because DPLPA satisfies differential privacy and minimizes the distribution position error, DPLPA is the best.

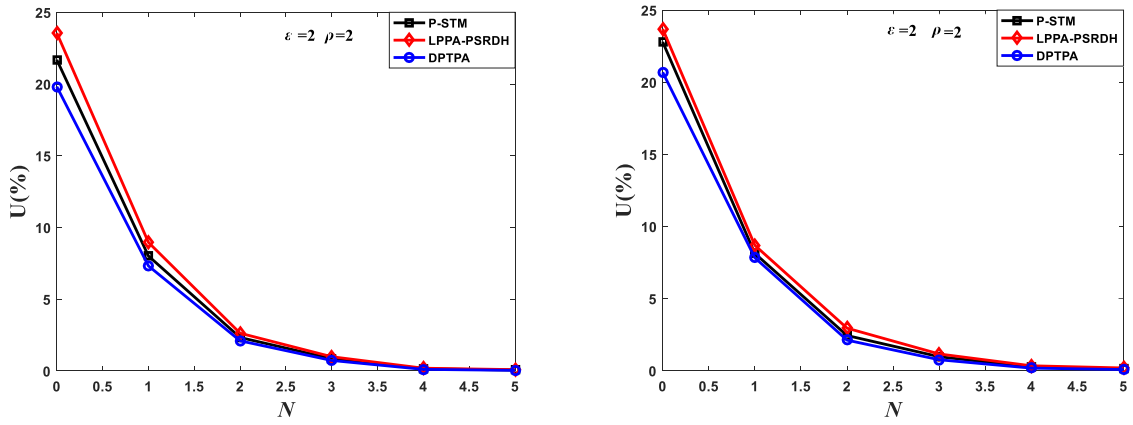
The influence of clustering density ρ on U is analyzed as shown in Fig. 11. U increases with the increase of ρ. Data utility of LPPA-PSRDU algorithm is the worst, and DPLPA is the best.

The influence of the number of protected locations N on U is analyzed as shown in Fig.12. U increases with the decrease of N. The data utility of DPLPA is the best, P-STM algorithm is the second.



(a) Geolife dataset (b) Gowalla dataset

Fig. 11 Influence of Cluster density ρ about three algorithms on U



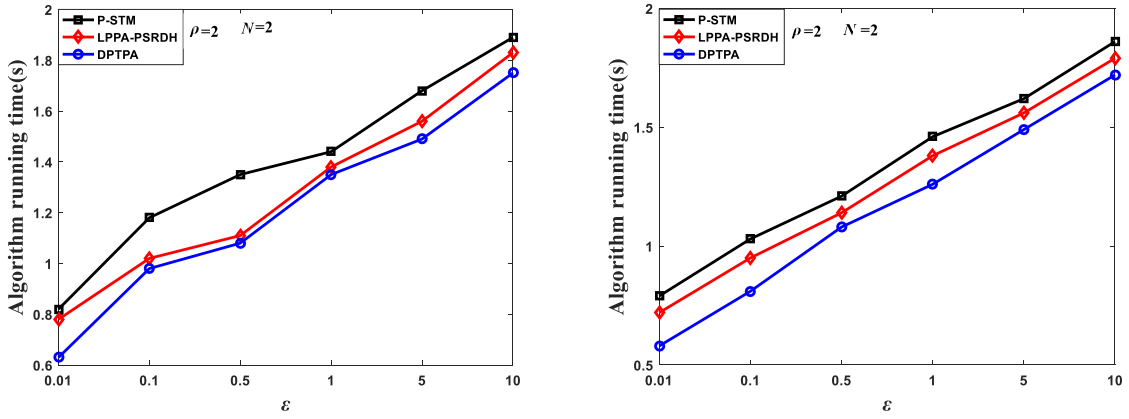
(a) Geolife dataset (b) Gowalla dataset

Fig. 12 Influence of the number of protection locations N about the three algorithms on the U

5.6 Algorithm time complexity

By running DPLPA, P-STM algorithm and LPPA-PSRDU algorithm, set the privacy budget ε, cluster density ρ and the number of locations N to be protected as different values, and comparing the operation efficiency of the three algorithms.

The influence of the privacy budget ε on the running time is analyzed as shown in Fig. 13. The running time of the algorithm increases with the increase of ε. The larger the privacy budget ε is, the longer it takes to allocate the privacy budget, the longer the algorithm runs. At the same time, because DPLPA algorithm only extracts the resident points and adds noise, the running time of DPLPA is the least, LPPA-PSRDU algorithm is the longest.

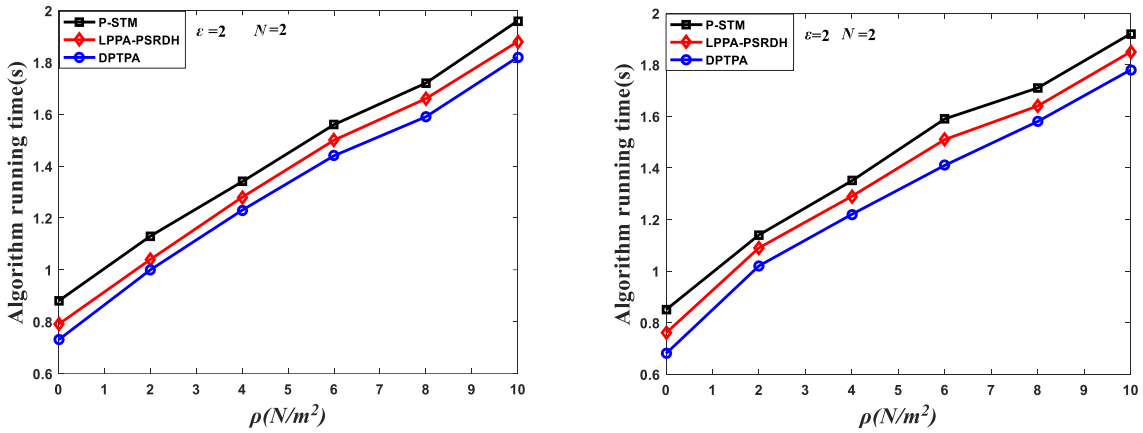


(a) Geolife dataset (b) Gowalla dataset

Fig. 13 Influence of the privacy budget ε about the three algorithms on running time

The influence of clustering density ρ about three algorithms on the algorithm running time is analyzed as show

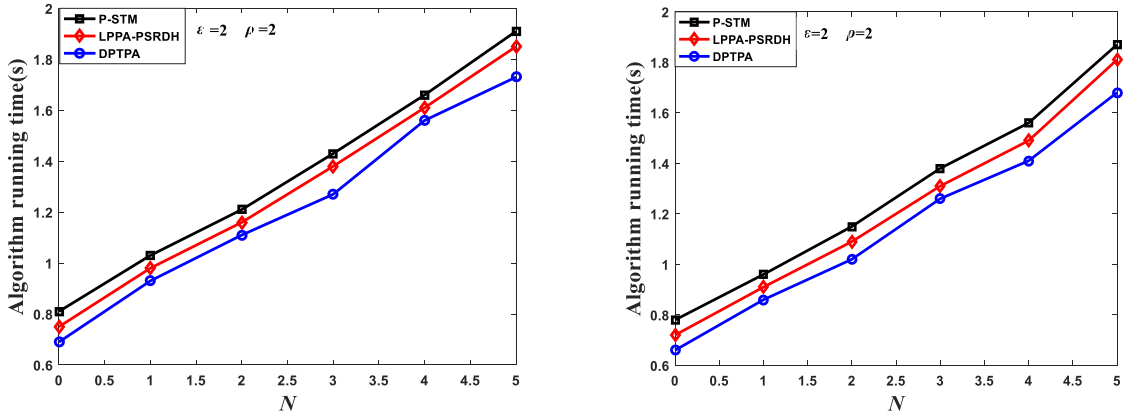
in Fig.14. Experiments show that on both datasets, the running time of the algorithm increases with the increase of the ρ. The higher the clustering density is the longer the running time is. The running time of LPPA-PSRDU algorithm is the longest on both datasets.



(a) Geoli dataset (b) Gowalla dataset

Fig. 14 Influence of clustering density ρ about three algorithms on running time

The influence of the number of protected locations N on the running time is analyzed as shown in Fig.15. Experiments show that on both datasets, the running time of the three algorithms increases with the increase of N. LPPA-PSRDU algorithm has the longest running time on both datasets, and DPLPA has the best running efficiency.



(a) Geolife dataset (b) Gowalla dataset

Fig. 15 Influence of N about the three algorithms on running time

6 Conclusions and future work

The privacy protection of continuous location data is studied in this work. Based on differential privacy mechanism, differential privacy location protection is realized by constructing interesting areas. In the continuous location

processing module, the interesting areas are constructed according to the users' access frequency to different locations. Then, L-cluster algorithm is proposed in the clustering module, which divides the continuous locations into different clusters according to the distance and density, and calculates the centroid of each cluster. In the location data protection module, DPLPA is proposed. Different resident points are selected and the privacy budget is allocated for them and the centroid. Laplace noise conforming to the differential privacy mechanism is added to protect the users' location privacy. A series of simulation experiments show that DPLPA has a good privacy protection efficacy, high data utility and low algorithm time consumption.

In this study, we only considered the continuous location protection under the users' historical state, and did not consider the real-time protection effect of data. Moreover, the data utility can be further improved. In the future work, corresponding data consistency algorithms can be proposed for different situations to solve the problem of data inconsistency caused by adding Laplace noise.

7 Statements and Declarations

7.1 Ethical Approval and Consent to participate

The experiments was approved by all participants.

7.2 Human and Animal Ethics

Not applicable.

7.3 Consent for publication

This article has not been published before and has not been considered for publication elsewhere. This publication will not be published in English or any other language without the consent of the author.

7.4 Availability of supporting data

The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

7.5 Competing interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

7.6 Funding

This study was supported by the National Natural Science Foundation of China under grant nos. 61301250 and

61702316; China Scholarship Council under Grant [2020]1417; Key Research and Development Project of Shanxi Province under Grant 201803D421035; Natural Science Foundation of Shanxi Province under grant no. 20210302123338.

7.7 Authors' contributions

Wang Bo and Li Hongtao conceived of the presented idea and wrote original draft preparation. Li Hongtao developed the theory and performed the computations. Wang Bo and Guo Yina verified the analytical methods. Guo Yina encouraged Wang Bo to investigate and supervised the findings of this work. Ren xiaoyu prepared figures and experiment data. All authors discussed the results and contributed to the final manuscript.

7.8 Acknowledgements

The authors would like to thank every teacher and postgraduate student who contributed to the background information of this study in the School of Electronic Information Engineering, Taiyuan University of Science and Technology, and College of Mathematics and Computer Science, Shanxi Normal University; In particular, the support of prof. Li Hongtao and Prof. Guo Yina during the all stages of this research is also appreciated. Finaly, We also would like to express our sincere gratitude to the editor and reviewers for their reading and appraising.

7.9 Authors' information

Bo Wang received the master’s degree in Ocean University of China in 2018, and she is currently pursuing the PhD degree with the Taiyuan University of Science and Technology. Her research interests include communication security, privacy preservation and federated learning.

Yina Guo and a PhD degree from Taiyuan University of Science and Technology in 2014. She is currently a Professor at Taiyuan University of Science and Technology. Her research interests include Brain-Computer Interfaces, blind signal processing and machine learning.

Hongtao Li received a PhD degree from Xidian University in 2015. He is currently a Professor at Shanxi Normal University, China. His research interests include IoT security and privacy preservation.

Xiaoyu Ren received the master’s degree in Shanxi Normal University in 2021. Her research interests include wireless network security and privacy preservation.

References

[1] Kang H Y, Zhu W X. Privacy preservation for location-based services [J]. Journal of Shandong University (Science Edition), 2018, 53 (11):35-50. http://en.cnki.com.cn/Article_en/CJFDTOTAL-SDDX201811005.htm

[2] Xu X Y, Chen H F, Xie L. A Location Privacy Preservation Method Based on Dummy Locations in Internet of Vehicles [J]. Applied Sciences, 2021, 11(10). http://dx.doi.org/10.3390/app11104594

[3] Jagdale B N, Bakal J W. A novel authentication and authorization scheme in P2P networking using location-based privacy[J]. Evolutionary Intelligence, 2020(2).

[4] Zhang R L, Zhao X H, Wu X N. Location privacy protection method against edge attacks and semantic attacks[J]. Computer application research, 2021, 38 (02):559-563.

[5] Dwork C, Kenthapadi K, Mcsherry F, et al. Our data, ourselves: Privacy via distributed noise generation[J]. Lecture Notes in Computer Science, 2006: 486-503. https://doi.org/10.1007/11761679_29

[6] Xiong J B, Ren J, Chen L, et al. Enhancing privacy and availability for data clustering in intelligent electrical se rvice of IoT, IEEE Internet of Things Journal, 2019, 6(2):1530-1540. https://doi.org/10.1109/JIOT.2018.2842773.

[7] Xiong J B, Ma R, Chen L, et al. A personalized privacy protection framework for mobile crowdsensing in IIoT, IEEE Transactions on Industrial Informatics, 2020, 16(6):4231-4241. https://doi.org/10.1109/TII.2019.2948068

[8] He J S, Du J H, Zhu N F. Research on *k*-anonymity Algorithm for Personalized Quasi-identifier Attributes [J]. Information network security, 2020, 20(10):19-26.

[9] Liu Q, Yu J, Han J, et al. Differentially private and utility-aware publication of trajectory data[J]. Expert Systems with Applications, 2021, 180(7):115120. https://doi.org/10.1016/j.eswa.2021.115120

[10] Song Cheng, Zhang Yadong, Wang Lei, et al. Research on *k*-anonymity privacy protection scheme based on bilinear pairings [J]. The Journal of China Universities of Posts and Telecommunications, 2018, 25(05):12-19. https://doi.org/10.19682/j.cnki.1005-8885.2018.0021

[11] Zhang Y B, Zhang Q Y, Li Z Y, et al. A *k*-anonymous Location Privacy Protection Method of Dummy Based on Geographical Semantics [J]. International Journal of Network Security, 2019, 21(6):937-946. http://dx.doi.org/10.1109/icpr.2014.715

[12] Tian C, Xu H, Lu T, et al. Semantic and Trade-off Aware Location Privacy Protection in Road Networks via Improved Multi-objective Particle Swarm Optimization [J]. IEEE Access, 2021, vol.9, pp.54264-54275.

https://doi.org/10.1109/ACCESS.2021.3071407

[13] Khorasany M, Dorri A, Razzaghi R, et al. Lightweight blockchain framework for location-aware peer-to-peer energy trading[J]. International Journal of Electrical Power & Energy Systems, 2021, 127:106610.

[14] Zheng Y, Luo J, Zhong T. Service Recommendation Middleware Based on Location Privacy Protection in VANET [J]. IEEE Access, 2020 vol. 8, pp. 12768-12783. http://dx.doi.org/10.1109/ACCESS.2020.2964422

[15] Zhao X, Pi D, Chen J. Novel trajectory privacy-preserving method based on prefix tree using differential privacy [J]. Knowledge-Based Systems, 2020, 198(5):105940. https://doi.org/10.1016/j.eswa.2020.113241

[16] Li W H, Li C, Geng Y. APS: Attribute-aware privacy-preserving scheme in location-based services [J]. Information Sciences, 2020, 527(1). https://doi.org/10.1016/j.ins.2019.02.025

[17] Yuan T A, Mmk A, Mar A, et al. A privacy preserving location service for cloud-of-things system-Science Dire ct [J]. Journal of Parallel and Distributed Computing, 2019, 123:215-222. https://doi.org/10.1016/j.jpdc.2018.09.005

[18] Partovi A, Zheng W, Jung T, et al. Ensuring Privacy in Location-Based Services: A Model-based Approach [J]. 2020. https://doi.org/10.48550/arXiv.2002.10055

[19] H Wu, Li M, H Zhang. Enabling Smart Anonymity Scheme for Security Collaborative Enhancement in Location-Based Services [J]. IEEE Access, 2019, 7:50031-50040. https://doi.org/10.1109/ACCESS.2019.2911107

[20] B.A. Sabarish, Karthi R, Kumar T G. Graph Similarity-based Hierarchical Clustering of Trajectory Data [J]. Procedia Computer Science, 2020, 171:32-41. https://doi.org/10.1016/j.procs.2020.04.004

[21] Chen R, Fung B C, Desai B C, et al. Differentially private transit data publication: a case study on the montreal transportation system[C]. Knowledge discovery and data mining, 2012: 213-221.

[22] Hu M, Zhang Y, Huang H. Personalized location privacy protection algorithm in crowd sensing networks [J]. Application Research of Computers, 2019. http://en.cnki.com.cn/Article_en/CJFDTotal-JSYJ201903059.htm

[23] Wang S, Nepal S, Sinnott R, et al. P-STM: Privacy-Protected Social Tie Mining of Individual Trajectories[C]// 2019 IEEE International Conference on Web Services (ICWS). IEEE, 2019.

[24] Peng Z, An J, Gui X, et al. Location Correlated Differential Privacy Protection Based on Mobile Feature Analysis [J]. IEEE Access, 2019, 7:54483-54496. https://doi.org/10.1109/ACCESS.2019.2912006

[25] Li Y H, Cao X, Yuan Y, et al. PrivSem : Protecting location privacy using semantic and differential privacy[J]. World Wide Web, 2019, 22(6):2407-2436. https://doi.org/10.1007/s11280-019-00682-0

[26] Zhang, Q., Zhang, X., Li, W., Li, S., & Li, X.. (2019). Design of poi recommendation algorithm based on differ ential privacy protection. Computer Applications and Software. http://en.cnki.com.cn/Article_en/CJFDTotal-JY RJ201909044.htm

[27] Dang B, Wang Y, Zhou J, et al. Transfer Collaborative Fuzzy Clustering in Distributed Peer-to-Peer Networks[J]. IEEE Transactions on Fuzzy Systems, 2020, PP(99):1-1.

[28] Nagargoje A, Kankar P K, Jain P K, et al. Development of the geometrical feature extraction tool using DBSCAN clustering for toolpath generation in incremental forming. 2021. https://doi.org/10.21203/rs.3.rs-340927/v1.

[29] Cao K Y, Sun Q M, Liu H L, et al. Social space keyword query based on semantic trajectory[J]. Neurocomputing, 2020, 428(4):340-351. https://doi.org/10.1016/j.neucom.2020.02.130

[30] Luo, H., Zhang, H., Long, S. *et al.* Enhancing frequent location privacy-preserving strategy based on geo-Indistinguishability. *Multimed Tools Appl* **80,** 21823–21841 (2021). https://doi.org/10.1007/s11042-021-10789-0

[31] Razi F F. A hybrid DEA-based K-means and invasive weed optimization for facility location problem [J]. Journal of Industrial Engineering International, 2019, 15:499-511. https://doi.org/10.1007/s40092-018-0283-5