Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070209
ISSN 2371-8838 Vol. 7 Num. 2

Protecting Privacy through Differential Privacy of Location Data

Wenbing Tang, Weiyuan Zhang*

School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan, China

**Corresponding author*

Keywords: Differential privacy, k - means clustering, location privacy, LBS

Abstract: Location-based services (LBS) are now used in many different industries thanks to the quick growth of mobile computing devices, but this also puts user security and privacy at risk. A location protection system that combines clustering and differential privacy is suggested to solve the issue of uploading and sharing user location information with outside parties. Firstly, the surrounding location points are sorted and divided according to the density of location information, and k-means clustering is used to generalize them; the cluster centers are noise-added by a planar Laplacian mechanism under the premise of satisfying geographic indistinguishability to obtain the perturbed position of each location point, and then location privacy is protected. The experimental results proved that the algorithm in this paper has higher data utilization under the premise of ensuring location privacy.

1. Introduction

The explosion of sensors and mobile devices has made users' lives easier by enabling them to go where they want to go. However, the processing and storage capabilities of these devices have led to the leakage of some private information about users, including the use of location-based services (LBS), which capture users' location information [1]. To access services, users submit precise location data to LBS, yet providing unprocessed location data directly results in the exposure of users' private information. For example, when ordering a takeaway, getting transport or meeting other users, they have to disclose their location to the LBS server, and this collected location information will potentially reveal some basic information about us, which can be used by advertisers to serve advertisements and by criminals to carry out criminal activities [2]. The leakage of some of the sensitive location information of the users can cause a lot of damage to the users. Current research has focused on protecting users' information security and establishing a safe and effective model.

There have been numerous research results on LBS privacy protection schemes both at home and abroad [3-6]. Based on bilinear pairing theory and k-anonymity, Song et al [7] suggested an enhanced privacy technique in which the optimal fake location is chosen based on location information. Subsequently, Zhang [8] proposed a new geosemantics-based location privacy preserving approach that also satisfies k-anonymity, in which a maximum and minimum distance multicentric clustering algorithm is used to construct candidate sets and generate virtual location result sets based on their semantic similarity. However, data distribution and background knowledge assaults significantly restrict the notions of L-diversity and k-anonymity, making it impossible to provide a high level of privacy protection. The primary concept of a location tree is to build a tree structure in accordance with predetermined rules, using differential privacy and prefix trees [9] to preserve the privacy of track data. Track segments were stored at the tree's nodes. In order to anticipate the next potential location based on each site's transfer probability, Markov models were mostly utilized to describe the temporal connection between users' real locations [10]. Tareq [11] proposed a density grid-based clustering method for online data streams, using a grid-based approach to reduce the number of calls to the distance function and thus improve the quality of clustering. Sabarish [12] proposed a graph-based model for representing trajectory data that uses edge- and vertex-based measures to compute the similarity between trajectories, and clusters and identifies similar trajectories based on paths, thereby providing privacy guarantees for location privacy.

In order to maximise the accuracy of query results and improve the efficiency of the algorithm, this paper proposes a location privacy algorithm that fuses clustering and differential privacy under the condition of satisfying differential privacy. By combining differential privacy with the k-means clustering algorithm, the centre of mass points of the clustered set are selected and processed using the planar Laplace mechanism to obtain the perturbed locations, and the original locations are queried using the perturbed locations instead.

1) By combining k-means clustering and differential privacy, this paper proposes a mechanism that can effectively protect user location privacy.

2) To verify that the algorithm proposed in this paper outperforms other algorithms and improves data availability, the efficiency, and effectiveness of the proposed scheme were tested using real data sets.

2. Definitions and Related Concepts

2.1. Differential Privacy

2.1.1. Definition of Differential Privacy

Definition 1 (ε-differential privacy) the algorithm M satisfies ε-differential privacy (ε- DP), where ε > 0, when and only when for any two adjacent datasets D and D′ have:

$$\forall O \subseteq \mathrm{Range}(M) \colon \Pr[M(D) \in O] \leq \mathrm{Exp}(\varepsilon) \times \Pr[M(D') \in O] \tag{1}$$

Where Range (M) denotes the set of all possible outputs of Algorithm M. We consider two datasets D and D′ as adjacent datasets, denoted as D ≈ D′ when and only when D = D′ + L or D′ = D + L, where D + L denotes the dataset obtained by adding a location point L to the dataset D.

2.1.2. Privacy Budget

The privacy budget is the name given to the parameter. Equation (1) demonstrates that a smaller will result in a probability distribution of query responses produced by method M over two adjacent datasets that is more comparable, making it more challenging for an attacker to ascertain if an element is present in the dataset. As a result, there will be greater privacy protection. In order to strike a compromise between privacy and the usefulness of the results, the value of is therefore typically paired with certain constraints.

2.1.3. Sensitivity

Sensitivity Δf is defined as the maximum L1 parametric distance between the outputs of the query mapping function on adjacent datasets for a given query mapping function$f$.

Definition 2 (Sensitivity) For any function $f: D \rightarrow \mathrm{R}^{\mathrm{D}}$:

$$\Delta f = \overset{max}{{}_{D,D'}} \left\| f(D) - f(D') \right\|_1 \tag{2}$$

$|| \cdot ||_1$ denotes the Manhattan distance or L1 parameterisation. Sensitivity is divided into global sensitivity, which is determined by function f, and local sensitivity, which is determined by both function f and the specific data in a given data set D.

2.2. Geographic Indistinguishability

When analyzing the privacy of data points at a particular location in geospatial terms, a geographically undifferentiated technique is applied. With this method, differential privacy is extended to geographic location data and the idea of neighboring records in differential privacy is changed into two geographically nearby sites. The possibility that two locations will provide the same query location is relatively high when two locations are close to one another. As a result, an attacker cannot pinpoint the user's precise location using the query location they have provided. This is explained in the paragraphs that follow.

$$\Pr[M(D) \in L] \leq e^{\varepsilon \mathrm{r}} \times \Pr[M(D') \in L] \tag{3}$$

Where the parameter ε denotes the privacy budget per unit of distance and the parameter εr denotes the privacy budget inside a circle of random radius. According to equation (4), the user's real position is safeguarded inside a circle of radius r by adding Laplace noise to the true location points to ensure geographical indistinguishability.

2.3. Implementation Mechanisms

Theorem 1(Laplace mechanism): Given a dataset D. Let the function $f: D \rightarrow \mathrm{R}^{\mathrm{d}}$have a sensitivity Δf, then the randomized algorithm:

$$M(D) {=} f(D) {+} \left(Lap_1(\lambda), \ldots, Lap_k(\lambda)\right) \tag{4}$$

Where $\lambda = \Delta f \,/\, \varepsilon$ is the scale parameter and $Lap(\,\lambda)$ is the additional Laplace noise. The privacy budget ε is inversely proportional to the noise variable's relationship to the sensitivity of the query functionΔf. Greater additional Laplace noise and higher privacy protection arise from smaller ε.

The implementation of differential privacy measures in one-dimensional space uses the Laplace mechanism from Theorem 1. It must be extended to the continuous plane in order to attain geographical indistinguishability in two dimensions. The process calculates perturbation locations from a two-dimensional Laplace distribution centered on the real location given parameters and true locations. The planar Laplace distribution centered on the Cartesian coordinate system is converted into a polar coordinate form centered on the origin to speed up the calculation. The transformed probability density function was displayed in Equation (4).

$$P_\varepsilon(r, \theta) = \frac{\varepsilon^2}{2\pi} r e^{-\varepsilon r} \tag{5}$$

The perturbed position l'can be represented by(r, θ), where r is the distance between l andl', and θ is the angle between r and the baseline of the Cartesian coordinate system. Also$r =$ $C_\varepsilon^{-1}(p) = -\frac{1}{\varepsilon}(W_{-1}\left(\frac{p-1}{\varepsilon}\right) + 1)$,$\theta \sim U(0, 2\pi)$, if the perturbed position l' can be expressed as

$$l' = l + (r \sin\theta, r\cos\theta) \tag{6}$$

2.4. K-means Clustering

K-means is the most popular unsupervised learning method for clustering [13]. Let $X = \{x_1, x_2 \ldots, x_n\}$ be the data set in the d-dimensional Euclidean space R^d. Let $a = \{a_1, a_2, \ldots, a_1\}$ be the c-clustering centre. Let $z = [z_{ik}]_{n\times c}$, where z_{ik} is a binary variable $(z_{ik} \in \{0,1\}$. The k-means objective function is $J(z, A) = \sum_{i=1}^{n}\sum_{k=1}^{c} z_{ik}\left|\left|x_i - a_k\right|\right|^2$. The k-means algorithm iterates through the necessary conditions to minimise the k-mean objective function $J(z, A)$ and update the equations for the clustering centres and membership relations, respectively, as follows.

$$a_k = \frac{\sum_{i=1}^{n} z_{ij}x_{ij}}{\sum_{i=1}^{n} z_{ij}} \tag{7}$$

$$z_{ik} = \begin{cases} 1, if \left|\left|x_i - a_k\right|\right|^2 = \min\limits_{1\le k\le c}\left|\left|x_i - a_k\right|\right|^2 \\ 0, \qquad\qquad otherwise \end{cases} \tag{8}$$

$\left|\left|x_i - a_k\right|\right|$ is the Euclidean distance between datapoint x_i and cluster centre a_k.

3. Algorithm Design

To address the problem of low data utilisation of location privacy protection mechanism, k-means clustering is added to the differential privacy location protection mechanism. The mechanism can well protect the privacy of individual locations and make the perturbed locations similar to the real locations through the k-means algorithm and differential privacy, thus improving the availability of location data. The basic idea is as follows: for each location point, the points of interest with a distance less than r are assigned to the clustering cluster in which the location is located. The algorithm uses the cluster centroids to represent the user's activity region within a certain distance, and other location points within the region are removed to avoid location redundancy. Finally, to further protect the user's privacy, the original location is replaced by the centroid.

3.1. Location Point Pre-processing Module

Algorithm 1 Pre-processing noise addition

```
Input: $L = \{l_1, l_2, \ldots, l_n\}\ (l_i = (s_t, t_i)),\ \varepsilon,\ k$
Output: $C'$
1. up_sort(L) // Sorting the data set in ascending order according to the density of location points
2. C=get_up_sort(L) // Select k objects as initial clustering centres C={c1,c2, ...,ck}
3. for i=1 to k do
4.     $s_i' = s_i + r_i\cos\theta_i$
5.     $t_i' = t_i + r_i\sin\theta_i$
6.     $c_i' = (s_i', t_i')$// Adding noise to the initial clustering centroid $c_i \in C$
7. end for
8. return $C' = \{c_1', c_2', \ldots, c_k'\}$
```

Algorithm 1 first sorts the data set by location density and selects the k densest data objects as centroids, followed by polar transformation of the location data and noise addition. In step 4, the random noise is determined by $r_i = C_\varepsilon^{-1}(n_i)$ with $\theta_i \sim U(0, 2\pi)$, where n_i is a random number uniformly distributed between [0,1] and $C_{\varepsilon}^{-1}(n_i)$ is the integral function of the probability density function $C_{\varepsilon,r}(n_i)$ over [0, n_i].

3.2. Location Data Clustering Module

Algorithm 2 A location-preserving approach fusing clustering and differential privacy (K-DP)

```
Input: $L = \{ l_1, l_2, \ldots, l_n \} ( l_i = ( s_t, t_i))$, $C'$, $\varepsilon$, $k$
Output: $L'$
1.  $S_j = Ø(1 \leq j \leq k)$ // Initializing a cluster collection
2.  for i=1 to n do
3.      for j=1 to k do
4.          $d[i]=argmin(distance( l_i, c_j' ))$// Calculate the distance of each sample point $l_i \in L$ to each centroid $c'$ and determine the nearest centroid to it
5.          $S_j=c_j \cup \{l_i\}$// Divide it to the nearest centroid $c_j$', forming the set of clusters $S_j$
6.      end for
7.  end for
8.  for i=1 to k do
9.      $sum_j' = \sum_{i \in S_j} x_i + n_j$ // Sum of points within a set
10.     $num_j' = |S_j| + n_j$// Number of points in the set
11.     $x_j'' = (s_j'', t_j'') = sum_j' / num_j'$// Update the set $S_j$ centroid $x_j$”
12. end for
13. repeat 2-12 until Clustering convergence//Clustering centre $C''=\{x_j''\}$ after adding noise
14. return $L'= C''$
```

Algorithm 2 is a k-mean clustering difference privacy algorithm for the set of location points $L = \{ l_1, l_2, \ldots, l_n \}$. The first steps 2-7 divide each sample point $l_i \in L$ into the nearest centroid c_2' based on its two-parametric distance to each centroid c', forming a set of k clusters $S_1, S_2, \ldots, S_k$. For the set $S_j (1 \leq j \leq k)$, steps 8-12 then calculate the sum of points in the set $\sum_{i \in S_j} x_i$, and the number $|S_j|$, add noise to it according to the function sensitivity Δf and privacy budget ε, respectively, and finally obtain the updated centroid x_j'' of the set S_j .where the random noise n_j satisfies $n_j \sim Lap(b)$ and has $b = \Delta f / \varepsilon$. Finally, the protected location data set $L' = C''$ is output by step 14, where the perturbed clustered centroids replace the original locations.

3.3. Algorithm Security Analysis

The algorithm in this paper adds noise to the clustered data that obeys the $Lap(\lambda)$ distribution of the Laplace distribution, so that the noise results satisfy the differential privacy constraint. The proof is as follows: the probability density function$Pr(u) = \frac{1}{2b} e^{\frac{|u|}{b}}$ of the Laplace mechanism is known, x and y represent two different locations, the probability density function of Prx is Fm (x, f, ε), and the probability density function of Pry is Fm (y, f, ε), and for a given output value Z, there are:

$$\frac{Pr_x(Z)}{Pr_y(Z)} = \prod_{i=1}^{k} \frac{e^{-\frac{\varepsilon|f(x)_i - Z_i|}{\Delta f}}}{e^{-\frac{\varepsilon|f(y)_i - Z_i|}{\Delta f}}} = \prod_{i=1}^{k} e^{\frac{\varepsilon|f(y)_i - Z_i| - \varepsilon|f(x)_i - Z_i|}{\Delta f}}$$

$$\leq \prod_{i=1}^{k} e^{\frac{\varepsilon(|f(y)_i| - |f(x)_i|)}{\Delta f}} = e^{\frac{\varepsilon||f(y)_i| - |f(x)_i||_1}{\Delta f}} \leq e^{\varepsilon} \tag{9}$$

As a result, the technique in this work meets the need for differential privacy as stated in the definition.

4. Example Analysis

4.1. Experimental Setup

The hardware environment used for the experiments was an 11th generation Intel(R) Core(TM) i5 2.40GHz with 16GB of RAM, implemented using the Python programming language and Windows 10 as the operating system platform. The dataset used was Gowalla, a social network that collects location data and allows users to share their location by checking in. The dataset uses a public API to collect information, has 196,591 locations and 950,327 domains, and collected 6,442,890 check-ins from users over the period from February 2009 to October 2010.

4.2. Experimental Analysis

The MSE was used to measure data availability, comparing data availability using the differential privacy algorithm alone and analysing the impact of the privacy budget ε on data usage, where a lower error represents higher data availability. The analysis of the effect of the privacy budget ε on the algorithm error is shown in Figure 1.

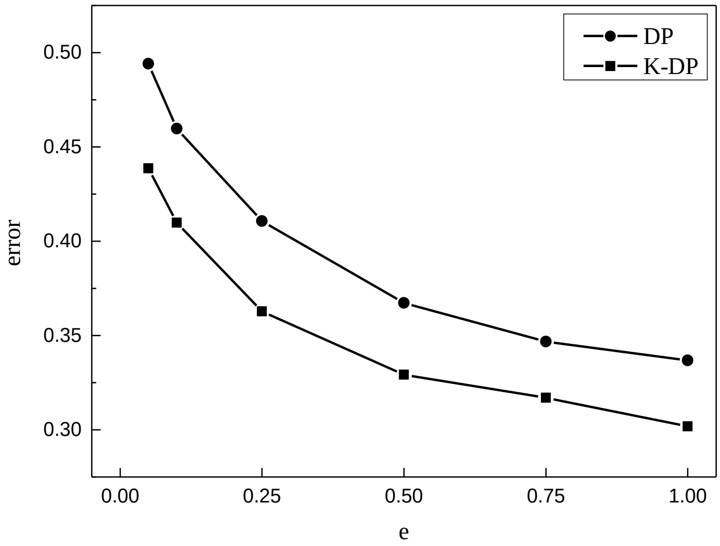
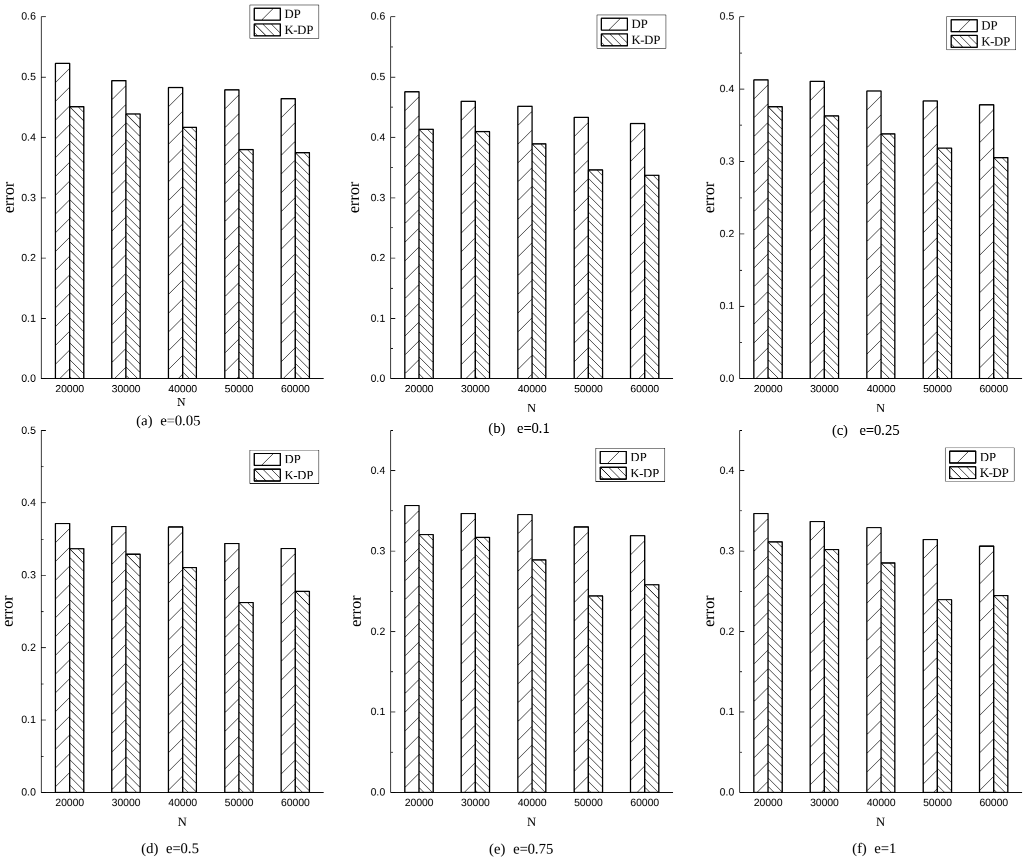


Figure 1: Error comparison of different algorithms under different privacy budgets.

From the Laplace probability density function, it can be deduced that the degree of privacy protection decreases with increasing ε. This experimental result also reflects this theory. Furthermore, it is easy to see from the experimental results that the privacy level of this algorithm is better than that of the traditional differential privacy algorithm because k-means clustering generates a centre of mass and all positions are replaced by a unique centre of mass, hence the better privacy level and higher data availability of K-DP.

The impact of N, the number of places to be secured, on the level of privacy protection is examined in Figure 2. The experimental findings show that as the number of sites to be protected N reduces, the degree of privacy protection improves; the greater N, the poorer the degree of privacy protection. Similar to how it can be observed that the data availability of the technique in this study is higher than that of conventional differential privacy approaches.



(a) e=0.05

(b) e=0.1

(c) e=0.25

(d) e=0.5

(e) e=0.75

(f) e=1

Figure 2: The influence of quantity N on the degree of privacy protection.

5. Conclusion

In this paper, a new solution to the privacy problem of location data is proposed, and an algorithm that fuses k-means clustering with differential privacy is designed to interfere with the user's location data. In order to combine the strong privacy of differential privacy and the advantage of small error of k-means clustering, a geographically indistinguishable model that satisfies the differential privacy is used for the noise addition to avoid the impact on the clustering results. Through the theoretical analysis and experiments, it is proved that the algorithm in this paper can improve the utilisation of the data and at the same time effectively protect the security of the user data.

References

[1] Jiang H.Y., Zeng J.Q., & Han K. (2021) Research on Location Privacy Protection Methods for Mobile Users in 5G Environment. Transactions of Beijing Institute of Technology, 41(01), 84-92.

[2] Yang Y., Wang R.C. (2020) Location based service location privacy protection method based on location security in augmented reality. Journal of Computer Applications, 40(05):1364-1368.

[3] Xiong, J., Ren, J., Chen, L., Yao, Z., Lin, M., Wu, D., & Niu, B. (2018). Enhancing privacy and availability for data clustering in intelligent electrical service of IoT. IEEE Internet of Things Journal, 6(2), 1530-1540.

[4] Xiong, J., Ma, R., Chen, L., Tian, Y., Li, Q., Liu, X., & Yao, Z. (2019). A personalized privacy protection framework for mobile crowdsensing in IIoT. IEEE Transactions on Industrial Informatics, 16(6), 4231-4241.

[5] He J., Du J., & Zhu N. (2020) Research on k-anonymity Algorithm for Personalized Quasi-identifier Attributes. Information network security, 20(10), 19-26.

[6] Liu, Q., Yu, J., Han, J., & Yao, X. (2021). Differentially private and utility-aware publication of trajectory data.

Expert Systems with Applications, 180, 115120.

[7] Cheng, S., Yadong, Z., Lei, W., & Zhizhong, L. (2018). Research on k-anonymity privacy protection scheme based on bilinear pairings. The Journal of China Universities of Posts and Telecommunications, 25(5), 18-25.

[8] Zhang, Y. B., Zhang, Q. Y., Li, Z., Yan, Y., & Zhang, M. Y. (2019). A k-anonymous Location Privacy Protection Method of Dummy Based on Geographical Semantics. Int. J. Netw. Secur, 21(6), 937-946.

[9] Zhao, X., Pi, D., & Chen, J. (2020). Novel trajectory privacy-preserving method based on prefix tree using differential privacy. Knowledge-Based Systems, 198, 105940.

[10] Tian, Y., Kaleemullah, M. M., Rodhaan, M. A., Song, B., Al-Dhelaan, A., & Ma, T. (2019). A privacy preserving location service for cloud-of-things system. Journal of Parallel and Distributed Computing, 123, 215-222.

[11] Tareq, M., Sundararajan, E. A., Mohd, M., & Sani, N. S. (2020). Online clustering of evolving data streams using a density grid-based method. IEEE Access, 8, 166472-166490.

[12] Sabarish, B. A., Karthi, R., & Kumar, T. G. (2020). Graph similarity-based hierarchical clustering of trajectory data. Procedia Computer Science, 171, 32-41.

[13] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.