Design and Application of Teaching Case of Big Data Analysis Based on Differential Privacy

Wenju Liu, Huicong Jiao, Ze Wang

Abstract— **Aiming at the problem of poor data availability caused by excessive adding noise during trajectory privacy protection by differential privacy, combined with the teaching experience in big data privacy protection in recent years, a trajectory privacy protection method based on semantic privacy level was proposed. The case design and teaching method of trajectory privacy protection are discussed in order to provide reference for the applied teaching reform of differential privacy.**

Index Terms—**Location semantics, differential privacy, ladder mechanisms, privacy budgets**

I. INTRODUCTION

The advent of the era of big data is accompanied by the emergence of a large number of intelligent electronic devices, such as the popularity of many lightweight electronic devices such as wristbands and wireless earphones, providing great prospects for the development of LBS location services. Sport bracelet and a smart phone GPS has been widely used, such as Meituan, alipay software on all need to support LBS location positioning tracking algorithm, with the development of the era of 5 g network, LBS applications have been widely exists in the daily life, location services in bring convenient at the same time also brings personal private information leakage hidden trouble. While the real-time location algorithm on mobile devices brings us all kinds of network convenience, the data background is also constantly collecting information for summary, so as to analyze and process the data. Some malicious attackers will carry out inferential attacks through the collection of user information, so as to obtain user addresses, private health conditions, and other sensitive information without authorization. The disclosure of these information will seriously harm the personal safety and property security of users. Therefore, how to prevent personal privacy disclosure when releasing data is a hotspot to solve at present. Due to the strong location correlation and spatio-temporal correlation of trajectory data, many researchers now pay more attention to the privacy protection of trajectory release, and studying how to protect trajectory privacy is of great significance to the protection of LBS location privacy.

Difference (DP) due to its strict privacy technology proven technology to protect the privacy of data, the technology is not limited to a malicious attacker has multidimensional background knowledge, in the field of location privacy protection is one of the more common in recent years, the emerging technologies of LBS privacy protection technology the implementation of the basic principle is to add noise data, Before releasing the data set collected by the user's location information, the information equipment firstly adds noise to the data set, so that adding or deleting a piece of data in the data set will not affect the whole data set query result. Compared with traditional methods, differential privacy technology has strict proof of privacy protection, which makes researchers focus on DP technology in recent years. However, the differential privacy technology has some obvious defects. Due to the addition of a large amount of noise, the availability of data is greatly reduced. Therefore, how to balance the availability and privacy level of DP technology is a hot issue to be solved at present.

Trajectory data can be regarded as data composed of multiple single locations, and trajectory privacy protection is an important research hotspot at present. At present, many research schemes are aimed at improving data availability. Literature [1] proposes a privacy protection mechanism RPTR based on differential privacy, which uses dynamic sampling to protect real-time track data of vehicles and adds filters to improve data availability. Literature [2] proposed a regional privacy algorithm CPL based on geographical topological relations, and defined a γ -privacy model combining privacy level and differential privacy budget. De-min hu et al. [3] is a combination of the Hilbert curve and k - anonymous technology is put forward based on the m tree difference privacy protection method, the method on the basis of conventional quadtree structure is improved, the location data stored in the smallest rectangle node, root and intermediate node location information, realize when the data into the new position, The improved quadtree can synchronize dynamic update more efficiently, and the experiment shows that this method greatly improves the operation efficiency. In reference [4-5], Wang Hao et al. designed a two-dimensional Laplacian noise satisfying differential privacy by considering the characteristics of multidimensional trajectory data. Gaussian noise is processed by a special filter to generate noise sequence superimposed on the original trajectory for publication. At the same time, LAN Wei et al. [6] proposed a privacy protection algorithm based on the stay area, mining the region of interest for the trajectory privacy, and then setting the threshold to add noise only to the area of frequent stay points, so as to simplify the trajectory data and improve data availability. In literature [7], MDL algorithm is used to simplify the whole track, prefix tree is used to store track segment information, and differential privacy protection is carried out for track segment count. However, since prefix tree structure assumes that data have many same prefixes, most position points in the track are scattered in practice, this method is not ideal. Literature [8] proposed two spatial attack models based on spatial sparseness and maximum running speed, and proposed an effective consistency processing method constrained by

Manuscript received January 17, 2022

maximum running speed of users. Quadtree and R-tree were used to publish data index of space and road network respectively. Literature [9] proposed the concept of implicit location and established a model of inferred leak mode to anonymize sensitive locations through location substitution and suppression. Meanwhile, two constraint models were proposed considering user behavior patterns and trajectory characteristics.

Aiming at the problem of poor data availability caused by excessive adding noise when differential privacy is used to protect trajectory privacy, a trajectory privacy protection method based on semantic privacy level is designed for teaching research.

II. Application Discussion of Teaching Cases

The teaching case is divided into several basic steps. First, the concept of differential privacy and its application in trajectory privacy protection are explained; second, the framework of the whole scheme is explained, questions are raised and solutions are designed; finally, experimental design is conducted to explain algorithm implementation and experimental design. This case caused a good response in the teaching discussion.

III. Design Idea

The goal of privacy protection research is to come up with techniques for modifying private data so that the modified data can be released securely (for study by third parties) without privacy attacks such as de-anonymization. Privacy disclosure and protection in trajectory data release are generally divided into two categories: first, the trajectory data set released contains only one trajectory. Each location point on the track corresponds to a record, and the user's privacy requirement is to ensure that the location at a certain point is secure. The second type is the published trajectory data set containing multiple trajectories. Every track is considered a record. The goal is to publish a sanitized set of data so that an attacker cannot know how the trajectory corresponds to the user. At the same time, the modified data should retain the overall information of the original data as much as possible under the premise of privacy protection, otherwise the published data will be worthless for research. Current research focuses on two aspects: on the one hand, what kind of protection can be provided by privacy protection technology, or what kind of attack can be resisted; On the other hand, how to preserve the useful information in the original data as much as possible while protecting privacy. In the use of differential privacy when the noise mechanism of location privacy protection, privacy protection level reflected in how much privacy budget distribution, privacy budget is smaller, the higher the level of privacy protection, how can the difference of privacy under the same level of privacy protection, add less noise, maintain data availability, is an important goal of the present study. Different for different users, the sensitive position, such as a doctor, the hospital is frequently accessed for his position, his sensitive position, high frequency access to the school, the user may be to think according to the frequency of visits to determine whether the location is sensitive position, also, for different users, privacy protection level requirements are different.

To solve above problems, this paper according to different user privacy level requirements of the differences between calculating trajectory for each user level of privacy protection area, carry different stay area contains the location of the semantic properties of the need for privacy protection for regional distribution of different privacy budget, making for regional privacy levels are high, can get a higher privacy protection.

The main design ideas of this paper are as follows:

(1) The stop-area mining algorithm obtains the set of stop-area in the whole trajectory, that is, the trajectory location region requiring privacy protection.

(2) for the trajectory of the area, the semantic properties of each position is used to calculate each region the average level of semantic privacy, according to different levels of non-uniform distribution of stay protected areas for each different privacy budget, on a single area, get first step generalized computing center positions, use ladder add noise noise mechanism for this position, The generalized position is disturbed by noise.

(3) Simulation experiments are carried out on real trajectory data sets, and the experiments show that the proposed scheme improves the availability of data and guarantees the privacy of data compared with the existing methods of evenly allocating privacy budget.

The basic framework of trajectory privacy protection method based on semantic privacy degree is shown in the figure.

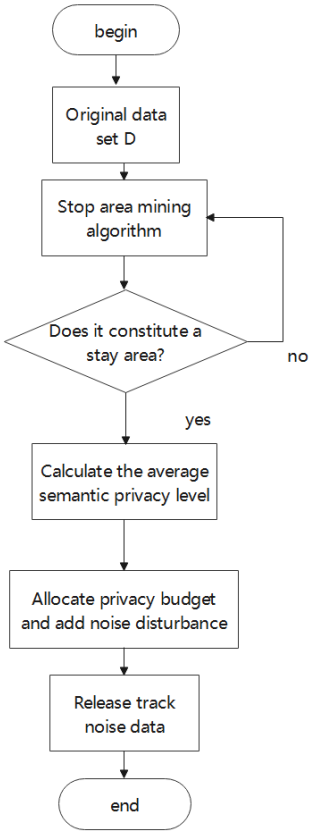


fig. 1.flow chart

IV. Case Design Algorithm Description

For A user track T, suppose there are n stay areas, and A stay area A contains m spatio-temporal position points of the track. Firstly, the semantic privacy of each position point is

calculated by information entropy, and the average privacy of the stay area is calculated according to the number of positions in A, as shown in Formula (1) :

$$D_A = \frac{\sum_{j=1}^{m} sen(sloc)}{m} \quad (1)$$

For this area, assuming that the total budget for the epsilon, privacy for stay with different levels of semantic privacy area, the average level of privacy areas of high need higher level of privacy, so we have to add these regions more noise, due to the size of the noise and privacy is inversely proportional to the budget, so design is assigned to the area the privacy of A budget for:

$$\varepsilon_A = \varepsilon \times \left(1 - D_A \Big/ \sum_{i=1}^{n} D_i\right) \quad (2)$$

The subtrajectories contained in this region are

$$T = \{(x_i, y_i, t_i), (x_{i+1}, y_{i+1}, t_{i+1}) \ldots\ldots\ldots (x_j, y_j, t_j)\} \quad (3)$$

The region is generalized in the first step, and the generalized position L is obtained.Step noise is added to the generalized location according to the allocated privacy budget to obtain the noised data.

V. CASE EXPERIMENT ANALYSIS

4.1 Experimental environment and data set

The experimental environment of this paper is Windows10, the programming language is Python, and the data set is Geolife, a real user track sampling data set, which is the GPS track data set collected by 182 users of Geolife project of Microsoft research Asia for more than 5 years. GPS tracks are represented by a series of timestamp points, each containing information about latitude, longitude and time. The data set contains 182 users, containing a total of 17621 path, this article on the original trajectory data set, by Scott maps API crawl each location point on the map correspond to the semantic information, using semantic type code as a symbol of different semantic, added to the track each location attribute, trajectory privacy as an attribute.

4.2 Measurement Standard

In the trajectory privacy protection of differential privacy, the main solution goal is how to improve data availability while ensuring the level of privacy. Therefore, this paper uses two metrics to analyze the experimental results. Respectively, data availability analysis and privacy protection degree analysis under different privacy budgets.

4.2.1 Data Availability [6]

For trajectory privacy protection, on the one hand, data privacy is ensured, and on the other hand, the released data set should have good availability, so as to facilitate data analysis and utilization. The data availability of trajectory is usually calculated by a distance function. In the scheme proposed in this paper, trajectory with different lengths should be compared. Therefore, this paper uses dynamic bending distance to measure data availability of trajectory before and after privacy protection. The larger the dynamic bending distance is, the worse the trajectory data availability is. On the contrary, the smaller the distance is, the better the availability will be even after trajectory privacy protection, which is of great help to data analysis and follow-up research.

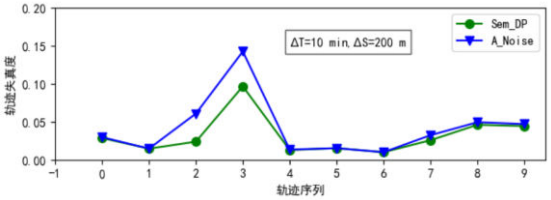
4.2.2 Degree of Privacy Protection [6]

Difference degree of privacy privacy related to privacy budget allocation, using the finite difference method the noise mechanism in the privacy to privacy protection of track, analysis the influence of different privacy budget on track availability to measure the degree of privacy protection of privacy difference method, the differential noise to give the definition of privacy, privacy is inversely proportional to the budget and availability, and is inversely proportional to the degree of privacy, the privacy budget is smaller, The higher the degree of privacy protection, the more noise added at the same time, the greater the impact on data distortion.

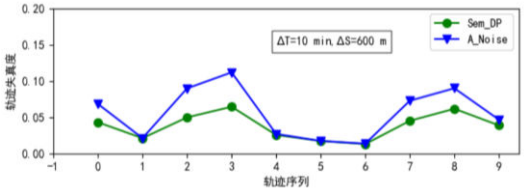
4.3 Experimental Results

The algorithm in this paper is named SenDP. For comparison, the algorithm proposed in literature [6] is named ANoise algorithm, which uses the method of evenly allocating privacy budget to protect the privacy of trajectory data.

(1) The influence of different distance parameter thresholds on the performance of the algorithm in this paper.



(a) The distance threshold is 200m



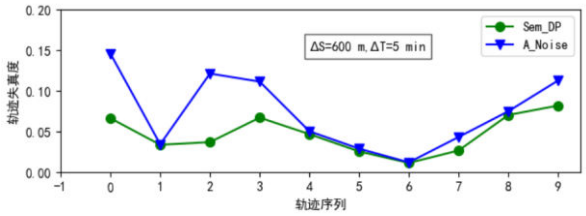
(b) The distance threshold is 600m

Fig 2. Distortion of trajectory at different distance thresholds

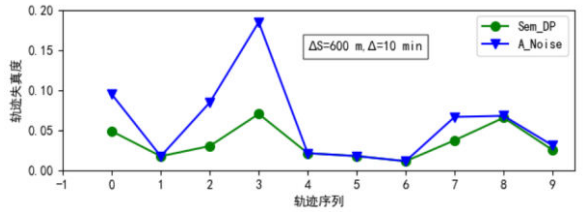
The influence of the proposed algorithm on data availability under different time thresholds and different distance thresholds is analyzed. FIG. 2 shows the distortion changes of tracks at different distance thresholds under the condition of fixed time threshold. In the figure, the distortion trend analysis is adopted under the fixed time of 10min and distance of 200m and 600m respectively. The whole graph is

divided into upper and lower two child graph, the total privacy budget we set to 10, the results in the figure shows that under the different distance threshold, under the condition of the same total privacy budget, SenDP distortion degree of the algorithm are less than ANoise algorithm, thus, under the same level of privacy protection, the proposed SenDP algorithm can add less noise, Improve data availability.

(2) The influence of different time parameter thresholds on the performance of the algorithm in this paper.



（a）The time threshold is 5 minutes



（b）The time threshold is 10 minutes

Fig 3. Distortion of trajectory at differenttimethresholds

(3) Privacy protection analysis.

The figure 4 shows that with the increase of privacy budget, track distortion degree, the lower the overall this is due to the differential noise mechanism of privacy, privacy, the larger the budget, to join the less noise, the less the distortion degree of trajectory, as you can see by the picture, under the same privacy budget, distortion degree path algorithm in this paper are less than contrast algorithm, with the increase of the budget of the privacy, The widening gap between the two indicates that the proposed algorithm achieves better usability when the privacy budget is large, while the proposed algorithm has little difference when the privacy budget is small. However, the proposed algorithm is still superior to the comparison algorithm.

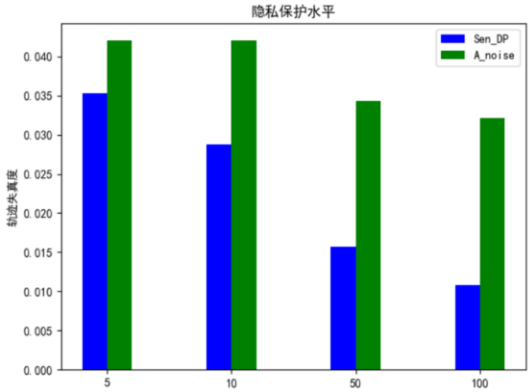


Fig 4. Usability comparisons under different privacy budgets

VI. CONCLUSION

Aiming at the problem of excessive noise of the differential privacy technology, the teaching is designed, and the mean value of the whole region is taken as the privacy rating of the whole region. According to different privacy levels, the total privacy budget is divided unevenly into each area to be protected. The teaching experiment shows the effectiveness of this method. It has good reference for future research.

REFERENCES

[1] Ma Z, Zhang T,Liu X, et al. Real-Time Privacy-Preserving Data Release Over Vehicle Trajectory[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8):8091-8102.

[2] Wu Y, Chen H, Zhao S. Differentially private trajectory protection based on spatial and temporal correlation[J]. Chinese Journal of Computers, 2018, 41(2): 309-322.

[3] Hu D M, LIAO Z J. Privacy protection method of differential privacy location in m-cross mean tree [J]. Microcomputer systems,2019,40(03):538-544.

[4] Hao W, XU Z, XIONG L, et al. CLM: differential privacy protection method for trajectory publishing[J]. Journal on Communications, 2017, 38(6): 85.

[5] Wang H,Xu Z.Differential privacy preserving method for trajectory clustering[J]. Huazhong Keji Daxue Xuebao (Ziran Kexue Ban)/Journal of Huazhong University of Science and Technology (Natural Science Edition), 2018, 46(1):32-36.

[6] LAN Wei, Lin Ying, Bao Lingyan, et al. Computer Science and Exploration, 2020, 14(1):14. (in Chinese)

[7] Hu demin, zhan han. Privacy protection method of user trajectory with predictable differential disturbance [J]. Journal of small microcomputer systems,2019,40(06):1286-1290

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

[9] Huo Zheng, Meng Xiaofeng. Journal of Computers, 2018, 41(2):13.