Location Big Data Partition and Publishing Method based on Sampling and Adjustment

Yan Yan, Lianxiu Zhang, Tao Feng, Pengshou Xie, Xin Gao

Abstract—**In order to maintain a trade-off between preserved privacy and enhanced utility of data publishing, a partition and publishing method for location big data is proposed based on sampling and adjustment. Firstly, the sampling with fixed time interval is used to simulate the publishing process of location big data, and differential processing method is designed to reduce the temporal and spatial redundancy of adjacent snapshots. Then, data update status at the current time is determined by the result of differential processing. Corresponding adjustment methods are designed for the grid-based and tree-based partition structure, and Laplace noise is added to the adjusted structure in order to realize differential privacy protection for the published data. Experiments show that the proposed partition and publishing method has larger advantages in improving regional query accuracy and the efficiency of algorithm.**

Index Terms—**location big data, privacy preserving, sampling, differential processing, differential privacy**

I. Introduction

THE rapid development of new technologies such as Mobile Internet, Cloud Computing, Internet of Things, Internet of Vehicles, etc. and the widespread popularity of intelligent devices have deepen the digitization of personal information and promoted the arrival of the era of big data. The amount of data has rapidly increased from GB, TB, to the level of PB and EB. The "2017-2023 China Big Data Industry Market In-depth Analysis and Development Trend Research Report" [1] shows that from 2011 to 2020, the amount of global data will increase from 1.8ZB to 44ZB dramatically. Big data contains immeasurable value and information. The collection, publishing, analysis, mining and application of big data have attracted the attention of governments, industries and research departments [2], [3], [4].

As the "natural entrance" of the Internet, location big data (LBD) is widely used in many hotspots such as intelligent transportation system (ITS), location based services (LBS), mobile social networking (MSN), etc. It has the characteristics of large volume, fast update speed, complexity and sparseness (low density). With the help of satellite positioning, network positioning and perceptual positioning, location information can be collected and released in real time, so that the public security support services can be provided, for example, the nearest police and ambulance can be dispatched to the accident site in time. Besides, with the help of geographic information system (GIS), location information can be used to help the public find out the state of traffic situation; planning a reasonable travel route and real-time navigation; achieve location-based information recommendation and advertising services; select the nearest stations, hotels, banks and other life service information.

However, location big data is closely related to personal privacy. Through the collection, reasoning, analysis and mining of location information, malicious attackers can not only obtain the location that users often stay, but also further predict their current position and future moving trajectory, leading to the disclosure of private information, living habits, health conditions, places of interest, consumption levels, etc., and may even endanger the property and life of users. Therefore, privacy protection of location big data is imminent.

Aiming at the privacy preserving data publishing problem of statistics location big data, many solutions have been proposed recent years. Privacy protection algorithm based on shape similarity of the trajectory position is proposed by adding the shape factor of trajectory and using the new trajectory similarity metric model [5], and data "mask" is formed by using the real original position information to maximize the similarity of the internal trajectories of the cluster while satisfying the trajectory k-anonymity. Privacy preference location protection method is proposed by combining the characteristics of k-anonymity and differential privacy [6], which not only can guarantee the maximum probability similarity of the anonymous set of points, but also realize the personalized setting of privacy protection level. The regional privacy level calculation algorithm [7] was proposed based on spatial geographic topological relationship, which analyzed the impact of published location on the real one before and after publishing. Meanwhile , the method also proposed a differential privacy location data publishing mechanismDPLRM, based on Markov probability transfer matrix. Considering the basic attributes associated with each location, an attribute-aware privacy protection scheme (APS) [8] was proposed to enhance the privacy of mobile users. Location-anonymity scheme was proposed [9] in order to solve the problem of location quality degradation, which use the occlusion area of the road movement model to maintain the anonymity of target node, thus reducing the tracking probability by attackers and improving the quality of service. Solution proposed in [10] ensures user to get the services they need from the service provider while protecting their private information. The partition and privacy preserving method of location big data divides the set of location information according to specific index structure, and releases statistics values within the index area to reduce the leakage risk of user's real location. By adding differential privacy

Manuscript received January 21, 2019; revised June 10, 2019. This work is supported in part by the Nature Science Foundation of China (61762059, 61762060, 61862040), China Scholarship Council (201808625040), the Nature Science Foundation of Gansu Province (18JR3RA156), and the Science and Technology Project of Lanzhou (2017-4-105).

Yan Yan is with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China (corresponding author e-mail: yanyan@ lut.edu.cn).

Lianxiu Zhang, Tao Feng, Pengshou Xie and Xin Gao are with Lanzhou University of Technology, Lanzhou, 730050, China (e-mail: zhang_lianxiu@163.com, fengt@lut.edu.cn, xiepsh_lut@163.com, gaoxin7@yeah.net)

noise to the statistics values, the effect of privacy protection can be further improved. Some typical partition methods such as tree-based partitioning method [11], [12], [13], grid-based partitioning method [14], and hybrid structure partitioning method [15], [16] are all aimed at reducing the non-uniform errors and noise errors generated during the partitioning process, as well as improving the availability of published data.

In this paper, we discuss the publishing process and privacy protection method of location big data. The main contributions of this paper are:

(1) The differential processing method is designed to reduce the temporal and spatial redundancy of adjacent snapshots after the sampling with a fixed time interval. It helps to reduce the amount of data to be processed, and improve the operating efficiency of the publishing algorithm.

(2) Different partition and adjustment methods are proposed for the grid-based and tree-based structures. These methods effectively balanced the noise error and non-uniformity error on the basis of location privacy protection, which will improve the query accuracy and data availability.

The rest of the paper is organized as follows: Section II introduce some basic knowledge about differential privacy and partition method. Sections III provide the basic principle and implement method of sampling and differential processing. Section IV presents the partition and adjustment algorithm for grid-based and tree-based structure. Section V shows a set of empirical studies and results. Section VI concludes the paper.

II. Basic Knowledge

Definition 1 (Differential privacy [17]): For all sibling datasets D and D' (differing on at most one element, written as $\|D - D'\| = 1$) and all $U \subseteq Range(R)$, a randomized algorithm R gives ϵ-ifferential privacy if :

$$P\left[R\left(D\right) \in U\right] \leq e^{\epsilon} \times P\left[R\left(D'\right) \in U\right] \tag{1}$$

$P\left(.\right)$ represents the risk probability of privacy disclosure. ϵ indicates the degree of privacy budget. The larger the ϵ, the lower the degree of protection of the user's privacy information, conversely, the higher the degree of protection. Mathematically, as long as e is small enough, it is difficult for an attacker to distinguish whether the query function acts on D or D' for the same output, thus achieving privacy protection.

Differential privacy protection can be achieved by adding an appropriate interference noise to the return value of the query function. However, adding too much noise can affect the availability of results, and too little can not provide sufficient security. Sensitivity is the key parameter that determines the amount of noise added. It refers to the maximum change caused by deleting any record in the dataset.

Definition 2 (Sensitivity): For the sibling datasets D and D', the sensitivity of the query Q' is the maximum value between the query results of D and D' :

$$S\left(Q\right) = max\left|Q\left(D\right) - Q\left(D'\right)\right| \tag{2}$$

$S\left(Q\right)$ is a parameter that determines the amount of noise added, indicating how sensitive the query function is to data changes.

Definition 3 (Laplace noise mechanism [18]): For any algorithm R, if the output result of R satisfies formula (3), algorithm R is said to satisfy the ϵ-differential privacy:

$$R\left(D'\right) = R\left(D\right) + Laplace\left(\frac{S\left(Q\right)}{\epsilon}\right) \tag{3}$$

Where $R\left(D'\right)$ is the noisy result of the algorithm acting on the published dataset, and $R\left(D\right)$ is the result on original dataset. The position parameter of Laplace distribution is 0, the scale parameter is $\frac{S(Q)}{\epsilon}$, the magnitude of noise is proportional to global sensitivity $S\left(Q\right)$, and inversely proportional to the privacy budget.

The noise error and non-uniformity error will be introduced into the published data during the process of spacial partitioning while using the tree-based or grid-based index structure. Therefore, various partition and publishing methods for location big data strives to improve the availability and quality of published data on the basis of reducing errors.

Definition 4 (Noise error): In order to prevent attackers from inferring user's specific location information through a large number of query results, Laplace mechanism or Exponential mechanism is often adopted to add noise to the statistical data, so as to satisfy the differential privacy protection requirement. The effect of the added noise on the query results is called the noise error:

$$Noise\_error\left(P\right) = \left|C\left(P\right) - C\left(P'\right)\right| \tag{4}$$

Where P represents the index area after partition, $C\left(P\right)$ and $C\left(P'\right)$ indicate the original statistical value and the statistical value after adding noise.

Definition 5 (Non-uniformity error): $P_i\left(i = 1, 2, ..., m\right)$ represents the partitioned areas intersect with the query area Q, which satisfies $\underset{1 \leq i,j \leq m \wedge i \neq j}{P_i \cap P_j} = \emptyset$. $r_i\left(i = 1, 2, ..., m\right)$ is the ratio of the query area Q to the current partition area. The non-uniformity error within the query region Q can be described as:

$$Non\_Uni\_error\left(Q\right) = \left|\sum_{i=1}^{m} r_i \cdot C\left(P_i\right) - C\left(Q\right)\right| \tag{5}$$

III. Data Publishing Via Differential Processing

Although the publishing process of location big data has the typical features of time series, it is not reasonable to process the statistical release of LBD in a data stream manner. On the one hand, partition and publishing of LBD provide statistical information about the number of users within a specific time or range. Taking the application demand of intelligent transportation system as an example, it is necessary to count the number of vehicles or users within the index area at regular intervals. Therefore, it is not necessary to process each new or exited location data in a data stream manner. On the other hand, real-time data processing implemented by streaming computing requires a large computational cost, which is unnecessary and not cost effective for LBD with highly redundant spatial distribution characteristics.

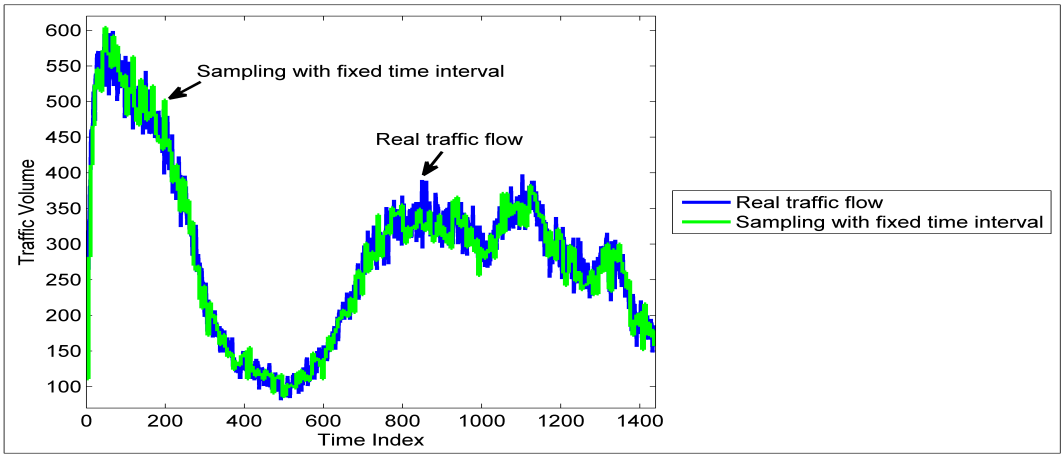


Fig. 1. Real-time traffic flow and sampling with fixed time interval

A. Location big data sampling with fixed time interval

The publishing process of LBD has the common features of big data such as large data size, dynamic changes, and uncertain update frequency. To achieve dynamic publishing of such LBD, an intuitive solution is to continuously publish snapshots of LBD and ensure that each release of the data meets the differential privacy protection requirements through a reasonable partitioning structure and added noise. Use t_i to indicate the sampling time, the published location big data can be represented as a set of snapshots $\{Y(t)\,|y(t) = X_{t_1}, X_{t_2}, ..., X_{t_i}, X_{t_{i+1},...}\}$, $i = 1, 2, ...$ The uniform sampling method has a fixed sampling time interval, it is simple and intuitive, and easy to be implemented by fast algorithm.

Fig.1 shows the real-time traffic flow (every 1 minute) of a city within one day and the result of uniform sampling (every 5 minute).Table I compares the original traffic flow data with the published sequence obtained by the uniform sampling method. The correlation between the original sequence and the sampling sequence is compared by using the expectation, variance, covariance and distortion rate as the metrics. Use Y to represent the expectation, variance and covariance value of the original sequence. $\widehat{Y}$ indicate the corresponding value of the sampling sequence; distortion rate can be defined as:

$$D = \frac{\left|Y - \widehat{Y}\right|}{Y} \tag{6}$$

It is not difficult to find from Fig.1 and Table 1 that the published data maintains the trend of the original data as a whole, and all the metrics reflect a high similarity and trend consistency between the original sequence and the sampling sequence.

B. Reduce redundancy through differential processing

Location big data publishing has the characteristics of real-time updating, large volume, and uncertain updating frequency. However, limited by time and traffic condition, location big data within a certain physical range does not change too drastically. In other words, the distribution characteristics of location big data have some redundancy in time and space. Therefore, differential processing of the location dataset at adjacent time can be performed to reduce the amount of data and improve the operating efficiency of the algorithm.

Use $Data = \langle D_1, D_2, ..., D_m \rangle$ to represents the set of location data, $D_i\,(i = 1, 2, ..., m)$ is the snapshot of dataset at every sampling point. For any location point (x, y) within the dataset, x represents longitude and y represents latitude.

Definition 6 (Differential processing): If the location point exists in the current snapshot D_i but not exist in the later snapshot D_{i+1}, then the point is marked as '-1', and the set of points is called the vanishing point set N_1, which can be specifically expressed as:

$$N_1 = D_i - D_{i+1} = \{(x, y)\,|\,(x, y) \in D_i and\,(x, y) \notin D_{i+1}\} \tag{7}$$

If the location point is not exists in the current snapshot D_i but does appeared in the later snapshot D_{i+1}, then the point is marked as '+1', and the set of points is called the new point set N2, which is specifically expressed as:

$$N_2 = D_{i+1} - D_i = \{(x, y)\,|\,(x, y) \notin D_i and\,(x, y) \in D_{i+1}\} \tag{8}$$

According to Definition 6, dataset of adjacent sampling time can be differential processed, and only the changed data points are left, which reduces the amount of processed data to some extent. Fig. 2 gives an example of the result before and after differential processing. The test dataset is selected from the New York TLC Trip Record Data[1]. L_1 and L_2 represent the current snapshot and the later snapshot, which contain 5000 and 15000 data points respectively.

IV. Partition Method Based on Adjustment

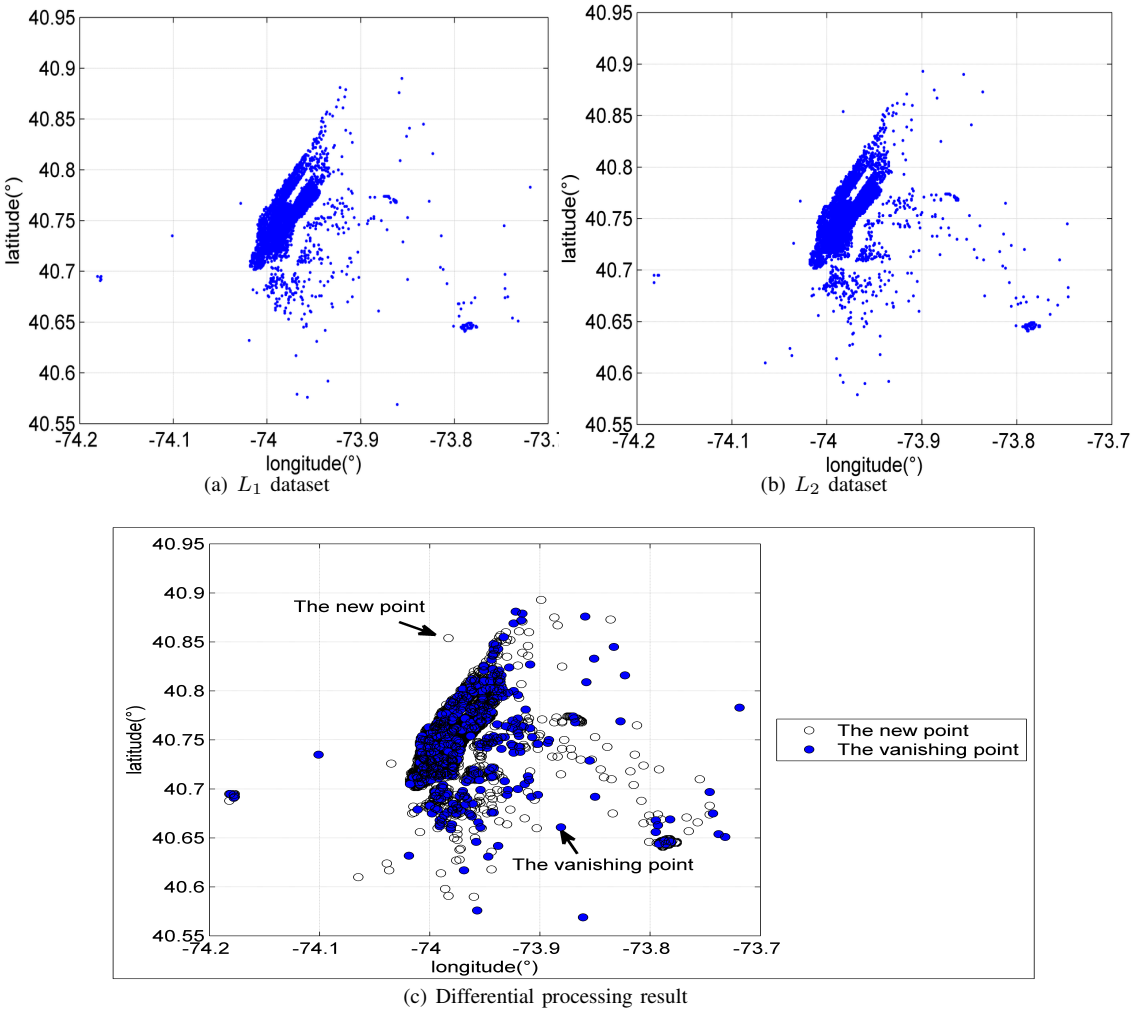
This section introduces the mechanism of partition and adjustment method for different index structure after differential processing, and then proposes corresponding partition and adjustment algorithms for the publishing of location big data.

As mentioned before, dynamically published location big data has certain redundancy in both spatial and temporal distribution. Therefore, for the two adjacent snapshots, their

[1] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

TABLE I. COMPARISON OF ORIGINAL SEQUENCE AND SAMPLING SEQUENCE

Parameter	Expectation		Variance		Covariance	
	E_x	Distortion rate of E_x	V	Distortion rate of V	Cov	Distortion rate of Cov
Original sequence	286.6706	-	14473.4676	-	14483.5326	-
Sampling sequence with fixed time interval	286.0799	0.0021	14526.7401	0.0037	14577.3560	0.0065



(a) L_1 dataset (b) L_2 dataset

(c) Differential processing result

Fig. 2. Example of differential processing

partition structure also has some similarities. With this similarity, we only need to make appropriate adjustments to the partition structure of the previous data snapshot to obtain the partition structure of the current data snapshot, and it is not necessary to execute the same partition strategy at all sampling moments. This will help speed up the execution of partitioning and publishing algorithms for location big data. In order to determine whether the data structure of the current snapshot needs to be adjusted, the degree of influence can be calculated on the vanishing point set N1 and the new point set N2 according to formula 8 and formula 9:

$$L_i = \frac{N_1^i + N_2^i}{\sum_{i=1}^{M}\left(N_1^i + N_2^i\right)} \tag{9}$$

$$L_0 = \frac{L_{max} + L_{min}}{2} \tag{10}$$

For each partitioned area *i*, N_1^i and N_2^i are the number of data points of vanishing point set and new point set. *M* represents the total number of areas in the snapshot. The larger the value of L_i, the greater the degree of influence, and the higher the possibility that the partition structure needs to be adjusted. If $L_i > L_0$, the partition structure needs to be adjusted; while if $L_i \leq L_0$, there is no need to adjust the partition structure.

A. Grid-based partition and adjustment method

For different partition structures, adjustment methods may be different. Take the grid-based partition structure for example, firstly the adaptive grid method (AG [18]) may be carried out on the first snapshot. The first layer partitions the data domain into $m_1 \times m_1$ grid cells of equal size, and obtains an noisy count N' for each cell by executing count query with a privacy budget $\alpha\epsilon$, where $0 < \alpha < 1$. The second layer partitions each cell into $m_2 \times m_2$ smaller grids which is adaptively chosen based on M':

$$m_1 = \sqrt{\frac{N\epsilon}{c_1}} \tag{11}$$

$$m_2 = \sqrt{\frac{N'(1-\alpha)\epsilon}{c_2}} \tag{12}$$

Where N is the number of data points, ϵ is the total privacy budget, c_1 and c_2 are some small constants depending on the dataset. In many cases, set $c_1 = 10$, $c_2 = \frac{c_1}{2}$.

Then, differential processing will be carried out on the adjacent two data snapshots, and the relationship between L_i and L_0 is judged according to formulas (9) and formulas (10) to determine whether the structure in the same area should be adjusted. The specific steps is shown in Algorithm 1.

Algorithm 1 Sampling and Adjustment based on AG (SA-AG)

Require: snapshot D_i, snapshot D_{i+1}, privacy budget ϵ_1
Ensure: Noisy Count N_{cell}
for snapshot D_i **do**
  partition D_i to get the first layer of cells region according to formula 10
  partition the first layer of cells according to formula 12
  differential processing of D_i and D_{i+1} according to definition 6
  get the number of data points in N_1 and N_2
  get L_i,L_0 according to formula 9, 10
  if $L_i > L_0$ **then**
    re-partition the region in D_{i+1} according to formula 11, 12
  else
    do not adjust the partition structure
  end if
end for
get the number of data point N_{real} for each cell region
add Laplace noise with a privacy budget ϵ_1
set $N_{cell} = N_{real} + Lap\left(\frac{1}{\epsilon_1}\right)$
return N_{cell}

B. Tree-based partition and adjustment method

The tree-based index structure partitioning method plays an important role in location big data partitioning and publishing technology. Whether a complete quad-tree structure or the improved methods based on it, the partition depth of the tree structure and the privacy budget allocation scheme should be adjusted to appropriately reduce the relative errors during the query and provide users with high quality services.

The space of location big data cannot be accurately divided into a single point, so generally two-dimensional partitioning tends to assume that the data distribution within an area is uniform. This estimation will introduce errors when the data points are not distributed uniformly. In fact, it is not optimal to partition different densities of spatial regions with the same standard. For sparse regions, this might result in over-partitioning, creating lots of empty areas, result in the increasing of noise error with little reduction in the non-uniformity error. On the other hand, for dense regions, this might result in under-partitioning and result in the increase of non-uniformity error. The inspiration of this phenomenon is that uniformity can be used as a standard for judging whether a region needs to be partitioned. Therefore, we redefined the metric amortization entropy to be the criterion for the re-partition of tree-based index structure.

Definition 7 (Data apportionment entropy [19]): a criterion that describes the non-uniformity of a node area to determine whether it is further divided. The specific definition is as follows:

$$DAE_i = -\sum_{i=1}^{M} \frac{c_i}{C} \log_2 \frac{c_i}{C} \tag{13}$$

$$DAE_0 = \frac{DAE_{max} + DAE_{min}}{2} \tag{14}$$

Where M represents the number of partition areas, C is the total number of data points, C_i is the count value for each partitioned area i.

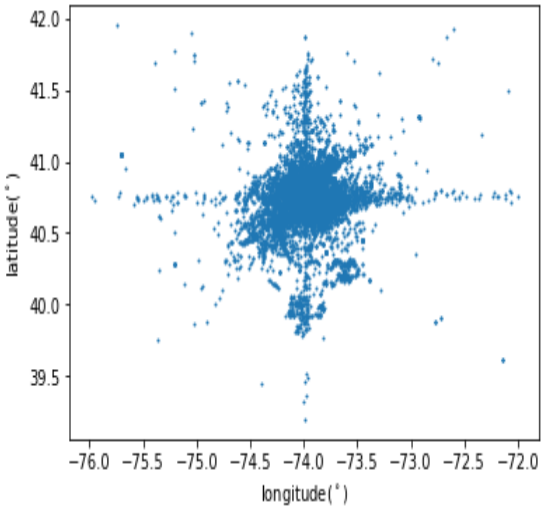
The larger the value of DAE_i, the greater the non-uniformity of the region, and the higher the possibility that the partition structure needs to be adjusted. So, for tree-based structure, firstly the Quad-post partition method [20] can be carried out on the first snapshot, and then the differential processing will be carried out on the adjacent two data snapshots. Finally, the relationship of DAE_i and DAE_0 can be compared according to formulas (13) and formulas (14) to determine whether the structure in the same area should be adjusted. Specific steps is shown in Algorithm 2.

Algorithm 2 Sampling and Adjustment based on Tree (SA-Tree)

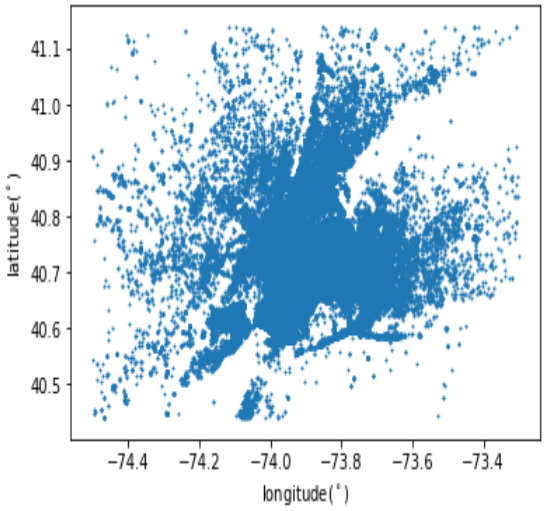
Require: snapshot D_i, snapshot D_{i+1}, privacy budget ϵ_2
Ensure: Noisy Count N_{tree}
for snapshot D_i **do**
  partition the spacial region according to Quad-post method
  differential processing of D_i and D_{i+1} according to definition 6
  get the number of data points in N_1 and N_2
  get L_i,L_0 according to formula 9, 10
  get DAE_i, DAE_0 according to formula 13 , 14
  if $L_i \leq L_0$ **then**
    do not adjust the partition structure
  else if $L_i > L_0$ **then**
    if $DAE_i > DAE_0$ **then**
      divide the region according to Quad-post method
    else
      do not divide the region
    end if
  end if
end for
get the number of data point N_{real} for each node region
add Laplace noise with a privacy budget ϵ_2
set $N_{tree} = N_{real} + Lapl\left(\frac{1}{\epsilon_2}\right)$
return N_{tree}

V. Experiental Result And Analysis

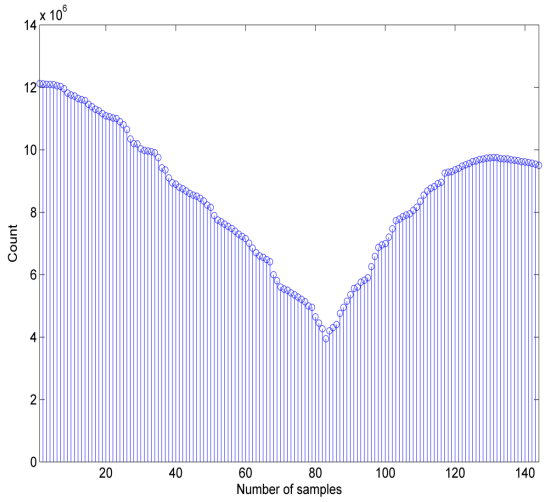
In order to verify the effect of partition and publishing method based on sampling and adjustment, we compares



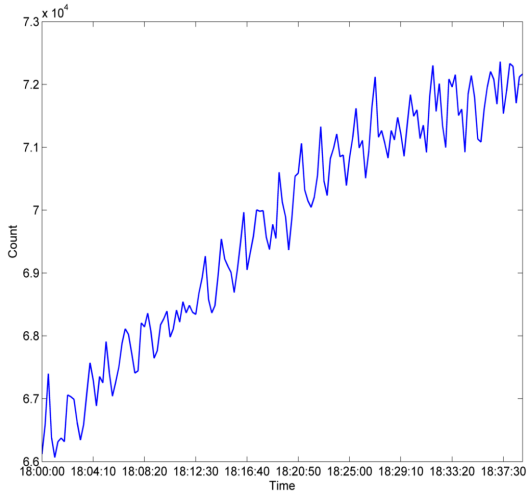
(a) Distribution of synthetic dataset



(b) Distribution of real dataset



(c) Synthetic sample sets



(d) Real sample sets

Fig. 3. Distribution status and sample size of experimental datasets

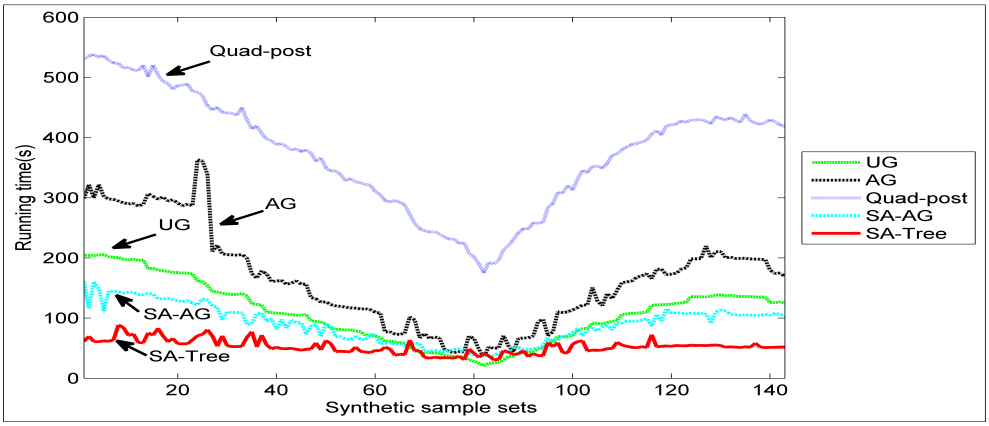
the efficiency and accuracy of the proposed algorithms with some existing methods, such as UG, AG and Quad-post. Experimental dataset contains the original data from the New York TLC Trip Record Data and the test dataset synthesized by it. The experimental platform uses Alibaba Cloud Server ECS (ecs.r5.xlarge: 4-core CPU, 32GB RAM, 100G SSD cloud disk, Windows Server 2008 R2 Enterprise Edition), and the algorithm is programmed by MATLAB R2015b software.

A. Analysis of efficiency

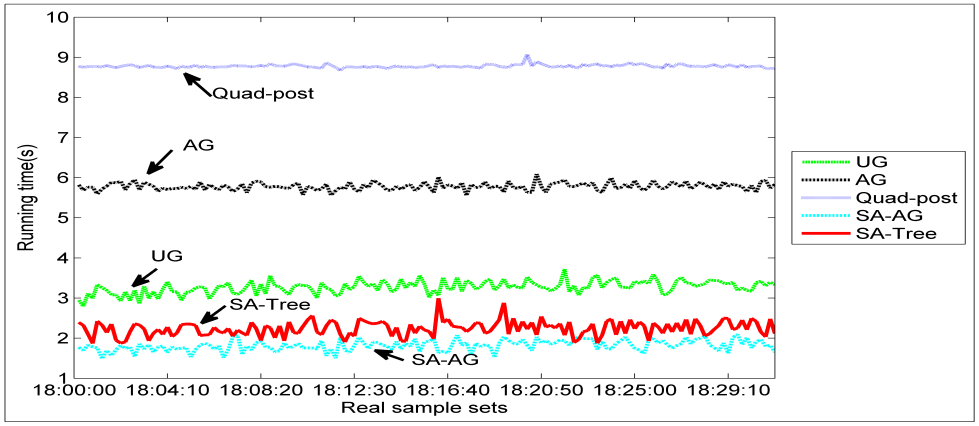
The operation efficiency of location big data partitioning and publishing algorithm mainly compares the overall time of building index structure and adding differential privacy noise to get the final published data. The performance of this part directly affects the feasibility of the algorithm in the dynamic big data publishing application environment.

This part of experiment was performed on real location big datasets and synthetic datasets. 150 sets of actual location big data were selected from the 2015 New York TLC Trip Record Data, which were obtained according to the sampling interval of 15s. 144 sets of synthetic big data were constructed based on the 2013 New York TLC Trip Record Data, simulating the dynamic publishing process of location big data. Fig. 3 shows the distribution status and sample size of the actual and synthetic datasets. The actual dataset is sampled from 18:00 to 18:38, and the change in the number of samples reflects the movement state of people during this time. The sample points of the synthetic dataset are selected from the same range, and the sample size is between 3.95 million to 12.12 million. The variation of the sample size simulates the location change process of the urban population within a certain period of time.

Fig. 4 compares the running time of UG, AG, Quad-post algorithms with the proposed SA-AG and SA-Tree algorithms on different experimental dataset. For the synthetic dataset, UG algorithm has the constant $c = 1000$. AG algorithm and SA-AG algorithm use the same privacy budget allocation ratio $\alpha = \beta = 0.5$. SA-Tree algorithm and Quad-post algorithm use the same partition depth $h = 6$ and total privacy budget $\epsilon = 0.1$. For the real dataset, the amount of data points after evenly spaced sampling is not very large, so the constant for UG algorithm is $c = 10$, and



(a) Running time of synthetic dataset



(b) Running time of real dataset

Fig. 4. Comparison of efficiency

other parameters are the same as above.

As can be seen from Fig. 3, the operating efficiency of various algorithms is basically proportional to the amount of data. The UG algorithm is a typical data independent partitioning method. Its partitioning process of the grid structure does not depend on the specific distribution state of data, so the input data only needs to be scanned once. Based on the first layer of UG partitioning, the AG algorithm performs adaptive partitioning again according to the grid density. Its execution process requires two scans of the input data, and the running time is slightly larger than the UG algorithm. The partition process of SA-AG algorithm only needs to locally adjust the partition structure of the data snapshot at the previous moment according to the data after differential processing, without rescanning all the data, so the SA-AG algorithm takes less time than the AG algorithm. Compared with the grid structure, partitioning method based on tree structure takes longer running time. The Quad-post algorithm needs to iteratively partition the two-dimensional space to form a complete quad-tree index structure. The SA-Tree algorithm improves the partitioning process of the Quad-post algorithm, determines the uniformity of the data distribution after differential processing to adjust the partition structure, avoids unnecessary partitioning process, and effectively improves the operating efficiency of the algorithm.

B. Analysis of accuracy

Partition and publishing algorithm for location big data mainly provides information inquiry services such as the number of users and traffic flow within a certain geographical range. The accuracy of range count query mainly compares the relative error of the published data with the original data within the query range Q. In this section, testing sets of location big data are selected from the synthetic dataset and the actual dataset, and six query areas with different range sizes are set according to the method of [18] (as shown in Table II). Relative error is defined as follows:

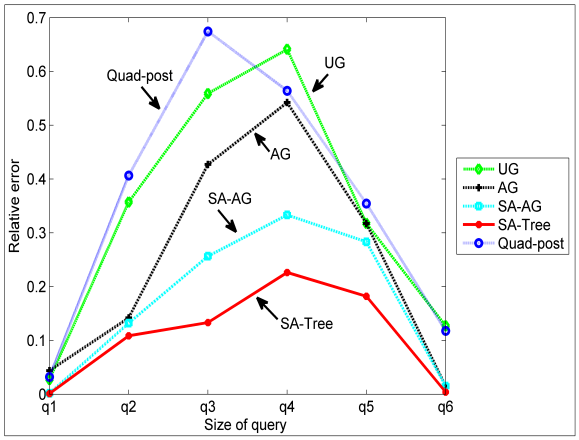
$$RE\_error = \frac{|Q(D) - Q(D')|}{max\{Q(D'), \rho\}} \tag{15}$$

Where $Q(D)$ is the correct answer to the query, $Q(D')$ is the noisy count, $\rho = 0.001 \times |D|$, $|D|$ represents the size of the dataset.

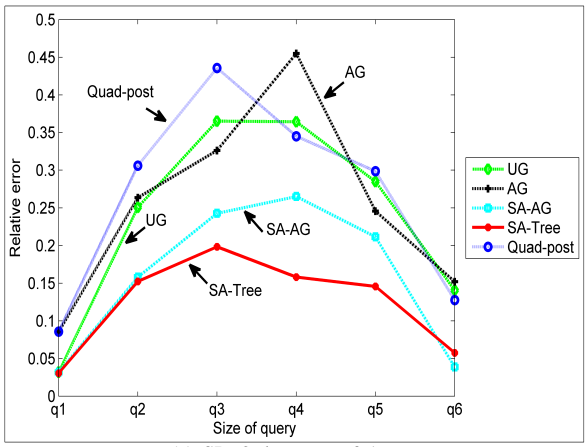
Fig. 5 to Fig. 8 show the relative error of range query of the mentioned algorithms under different datasets and different privacy budgets. Among them, the testing location big data of Fig. 5 and Fig. 6 is selected from the synthetic dataset, and the testing location big data of Fig .7 and Fig. 8 is selected from the real dataset. Comparing the experiment results it is not difficult to find that on the same dataset, relative errors of various algorithms are gradually reduced as the privacy budget increases. This is because, with the same sensitivity,

TABLE II. INFORMATION OF EXPERIMENTAL DATASET AND QUERY RANGE

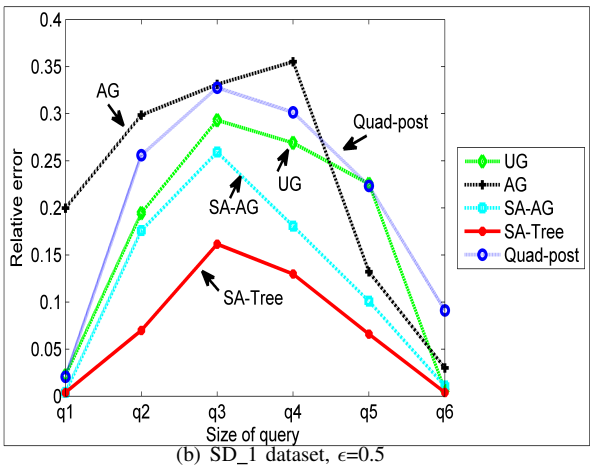
Parameter	Datasets			
	SD_1	SD_2	RD_1	RD_2
Total data amount	12096523	8920124	72364	66067
Coverage	1.4 × 0.9		1.2×0.7	
q1	0.01875×0.009375		0.015625×0.00625	
q2	0.0375×0.01875		0.03125×0.0125	
q3	0.075×0.0375		0.0625×0.025	
q4	0.15×0.075		0.125×0.05	
q5	0.3×0.15		0.25×0.1	
q6	0.6×0.3		0.5×0.2	



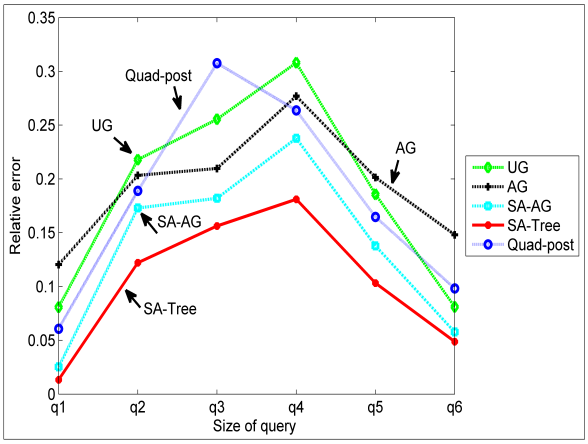
(a) SD_1 dataset, ϵ=0.1



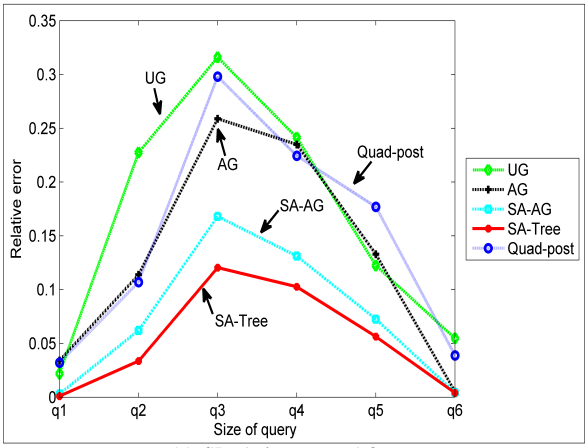
(a) SD_2 dataset, ϵ=0.1



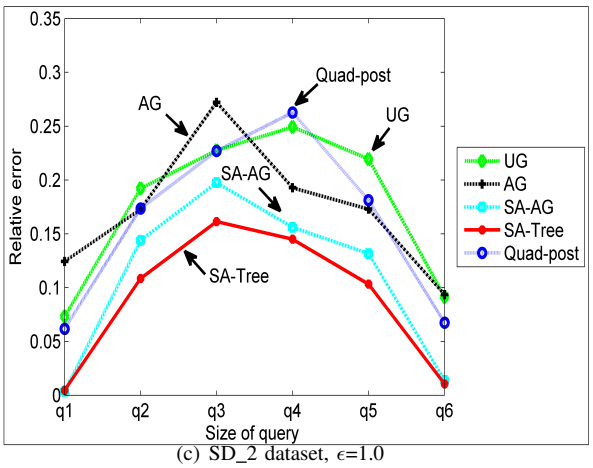
(b) SD_1 dataset, ϵ=0.5



(b) SD_2 dataset, ϵ=0.5



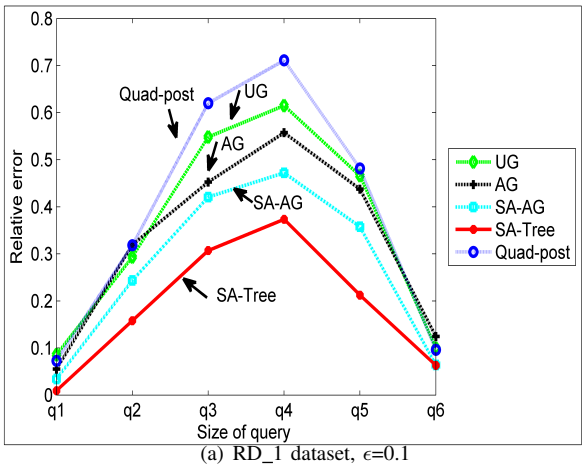
(c) SD_1 dataset, ϵ=1.0



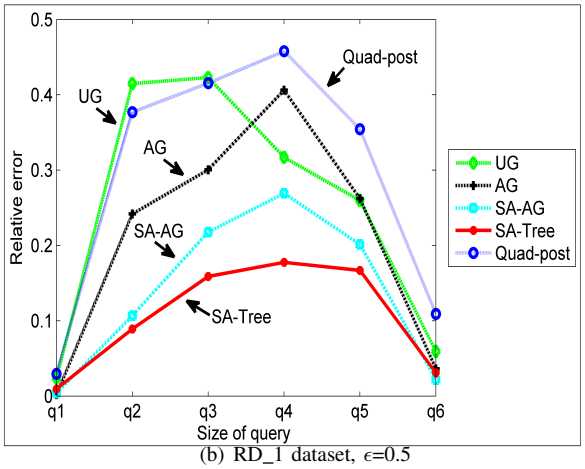
(c) SD_2 dataset, ϵ=1.0

Fig. 5. Comparison of query accuracy on SD_1 dataset

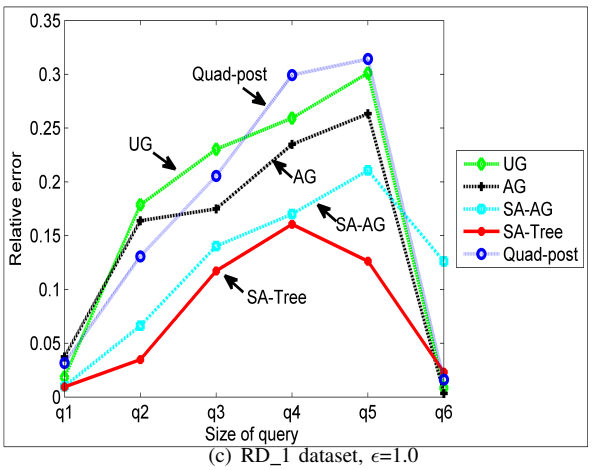
Fig. 6. Comparison of query accuracy on SD_2 dataset



(a) RD_1 dataset, ϵ=0.1

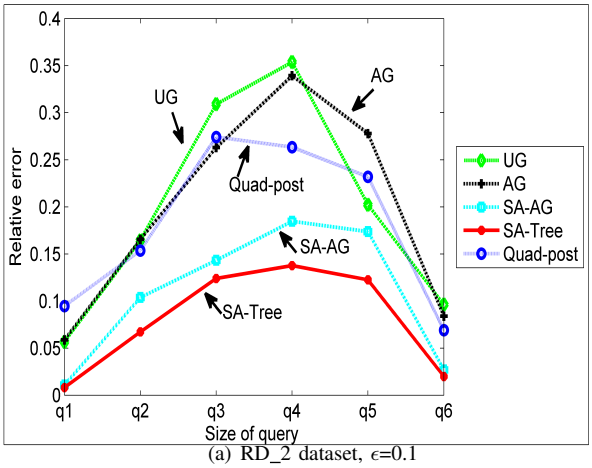


(b) RD_1 dataset, ϵ=0.5

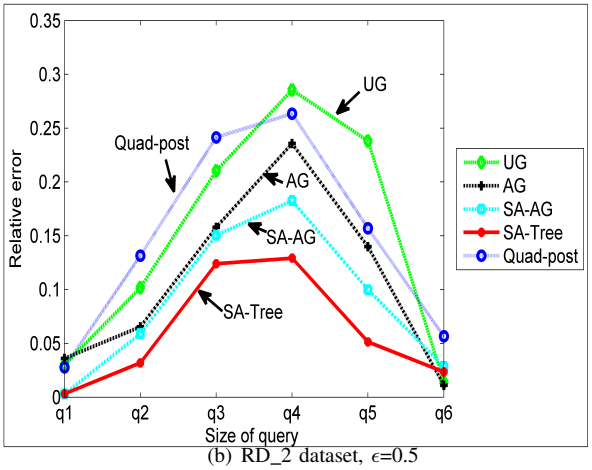


(c) RD_1 dataset, ϵ=1.0

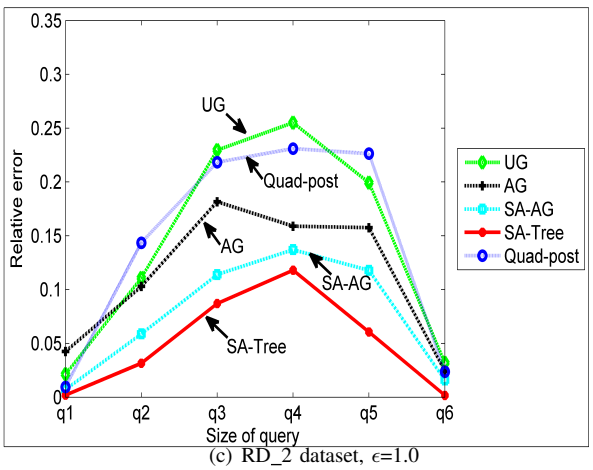
Fig. 7. Comparison of query accuracy on RD_1 dataset



(a) RD_2 dataset, ϵ=0.1



(b) RD_2 dataset, ϵ=0.5



(c) RD_2 dataset, ϵ=1.0

Fig. 8. Comparison of query accuracy on RD_2 dataset

the increase of differential privacy budget leads to a decrease of the added Laplace noise, so the deviation of the published result from the real data is reduced.

In the case of the same differential privacy budget, the UG algorithm has a larger non-uniform error and noise error, because it does not consider the actual distribution state of the location information. Therefore, the relative error of UG algorithm is large under various datasets and privacy budgets. The AG algorithm is improved on the basis of UG algorithm, it carried out an adaptive meshing process on the second layer to offset the error introduced by uniform partition on the first layer, which improves the query precision to some extent. The Quad-post algorithm based on quad-tree structure has certain advantages on large-scale query, but the performance is poor in the case of dense or uneven data distribution. SA-AG and SA-Tree algorithms are improved on the basis of AG and Quad-post algorithm, and the distribution characteristics of the data are considered more carefully. The problem of under-partition and over-partition can be avoided, and the relative errors under various datasets and privacy budgets are smaller. Therefore, the user's requirements for the accuracy of the query service can be better met.

VI. Conclusion

This paper mainly focus on the privacy protection and publishing issue of location big data. By analyzing the temporal and spatial correlation of location big data, a partition and publishing method is proposed based on sampling and adjustment. Dynamic location big data is sampled by uniform interval sampling, and adjacent data snapshots are differentially processed to obtain the updated data points. According to the different characteristics of grid-based and tree-based structure, corresponding partition adjustment algorithms are designed. Experimental results validate our methodology and show that our methods have better effect in query accuracy and operation efficiency than other methods.

References

[1] Zhiyan Consulting Group, "2017-2023 China Big Data Industry Market In-depth Analysis and Development Trend Research Report," http://www.chyxx.com/research/201708/548556.html.

[2] B. B. Raj, J. Frank, T. Mahalakshmi, "Secure Data Transfer through DNA Cryptography using Symmetric Algorithm," in *International Journal of Computer Applications 2016*, vol. 133, no. 2, pp. 0975-8887.

[3] A.S. Abad, H. Hamidi, "An Architecture for Security and Protection of Big Data," in *International Journal of Engineering(IJE) 2017*, Vol. 30, No. 10, pp. 1479-1486.

[4] Anupam Das, Shikhar Kumar Sarma, and Shrutimala Deka, "Data Security with DNA Cryptography," in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2019*, London, U.K., pp. 246-251.

[5] Chao Wang, Jing Yang and Jianpei Zhang, "Aparameterized location privacy protection method based on two-level Anonymity," *Journal of Communications 2015*, vol. 36, no. 2, pp. 148-161.

[6] Xiaodi Bi, Ying Liang, HongZhou Shi and Hui Tian, "Privacy preserving algorithm based on trajectory location and shape similarity," *Journal of Shandong University (Science Edition) 2017*, vol. 52, no. 5, pp. 75-84.

[7] Yuncheng Wu, Chen Hong, Suyun Zhao et al., "Differentially private trajectory protection based on spatial and temporal correlation," *Chinese Journal of Computers 2018*, vol. 41, no. 2, pp. 309-322.

[8] Shanthi, P. and Balasundaram, S.R, "A graph-based cloak algorithm to preserve location privacy in location-based services," *International Journal of Information Privacy, Security and Integrity (IJIPSI) 2015*, vol. 2, no. 2, pp. 138.

[9] Kenta Miura and Fumiaki Sato, "A hybrid method for user location anonymisation based on road mobility model," *International Journal of Adaptive and Innovative Systems (IJAIS) 2014*, vol. 2, no. 1, pp. 43-58.

[10] Xiaojuan Chen and Huiwen Deng, "A new scheme of preserving user privacy for location-based service," *International Journal of Electronic Security and Digital Forensics (IJESDF) 2018*, vol. 10, no. 4, pp. 417-433.

[11] Jun Wang, Rongbo Zhu, Shubo Liu and Zhaohui Cai, "Node Location Privacy Protection Based on Differentially Private Grids in Industrial Wireless Sensor Networks," *Sensors 2018*, vol. 18, no. 2, pp. 410-424.

[12] Yingjie Wu, Qing Lu, Jianping Cai and Xiaodong Wang, "Differential privacy two-dimensional data partitioning algorithm based on quadtree," *Journal of Huazhong University of Science and Technology (Natural Science Edition) 2016*, vol. 44, no. 3, pp. 410-424.

[13] Jun Wang, Shubo Liu, Yongkai Li, Hui Cao and Mengjun Liu, "Differentially Private Spatial Decompositions for Geospatial Point Data," *China Communications 2016*, vol. 13, no. 4, pp. 97-107.

[14] Qi Li, Yuqiang Li, Guicai Zeng and Aihua Liu, "Differential privacy data publishing method based on cell merging," *IEEE 14th International Conference on Networking, Sensing and Control (ICNSC) 2017*, pp. 778-782.

[15] Yan Yan, Xiaohong Hao and Lianxiu Zhang, "Hierarchical differential privacy hybrid decomposition algorithm for location big data," *Cluster Computing 2018*, no. 6, pp. 1-12.

[16] Yonghui Xiao, Xiong Li and Yuan Chun, "Differentially Private Data Release through Multidimensional Partitioning," *VLDB Conference on Secure Data Management 2010*, pp. 150-168.

[17] Lin Zhang, Yan Liu and Ruchuan Wang, "Differential Privacy Based Data Publishing Technology in Location Big Data Service," *Journal of Communications 2016*, vol. 37, no. 9, pp. 46-54.

[18] Wahbeh Qardaji, Weining Yang and Ninghui Li, "Differential Privacy Based Data Publishing Technology in Location Big Data Service," *Differentially Private Grids for Geospatial Data, Proceedings of the 29th International Conference on Data Engineering (ICDE) 2012*, New York, pp. 757-768.

[19] Jianyu Lu, Xiuqing Wang, Xuebin Wang et.al, "Metrics of distribution inhomogeneity during runoff and its application," *Journal of Hydroelectric Engineering 2015*, vol. 41, no. 11, pp. 24-28.

[20] Cormode G, Procopiuc M, Entong Shen and Srivastava D, "Differentially Private Spatial Decompositions," *Proceeding of 28th International Conference on Data Engineering (ICDE) 2011*, pp. 0-31.

Yan Yan received the Ph.D. degree in control theory and control engineering from Lanzhou University of Technology, China, in 2018. Her research interests include privacy preserving data publishing, differential privacy and information hiding. She is currently an Associate Professor at School of Computer and Communication, Lanzhou University of Technology, China. She is also an academic visiting scholar of Macquarie University from 2019 to 2020. She is a member of China Computer Federation.

Lianxiu Zhang is currently a master student of the School of Computer and Communication, Lanzhou University of Technology, China. She received the B.Eng. degree from the University of Tarim in 2017. Her research interests include network and information security and privacy preservation technology.

Tao Feng received the Ph.D. degree in computer architecture from Xidian University in 2008. He is currently a Full Professor and a Ph.D. Supervisor with the Lanzhou University of Technology. His main research interests include information security, provable theory of security protocols, wireless network security, and sensor network security. He is a member of the China Computer Federation and China Cryptography Federation.

Pengshou Xie is a professor and a supervisor of master student at Lanzhou University of Technology. His major research

interests include security of IoT, location privacy.

Xin Gao is currently a master student in School of Computer and Communication, Lanzhou University of Technology, China. She received the B.Eng. degree from Harbin Normal University in 2013. Her research interests include information security and dynamic clustering.