IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 10, 2019, accepted October 6, 2019, date of publication November 4, 2019,
date of current version November 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2951364

Dynamic Release of Big Location Data Based on Adaptive Sampling and Differential Privacy

YAN YAN[1,2], LIANXIU ZHANG[1], QUAN Z. SHENG[2], BINGQIAN WANG[1],
XIN GAO[1], AND YIMING CONG[1]

[1]School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
[2]Department of Computing, Faculty of Science and Engineering, Macquarie University, Sydney, NSW 2109, Australia

Corresponding author: Yan Yan (yanyan@lut.edu.cn)

This work was supported by the National Nature Science Foundation of China under Grant 61762059, Grant 61762060, and Grant 61862040.

ABSTRACT Data releasing is a key part bridging between the collection of big data and their applications. Traditional methods release the static version of dataset or publish the snapshot with a fixed sampling interval, which cannot meet the dynamic query requirements and query precision for big data. Moreover, the quality of published data cannot reflect the characteristics of the dynamic changes of big data, which often leads to subsequent data analysis and mining errors. This paper proposes an adaptive sampling mechanism and privacy protection method for the release of big location data. In order to reflect the dynamic change of data in time, we design an adaptive sampling mechanism based on the proportional-integral-derivative (PID) controller according to the temporal and spatial correlation of the location data. To ensure the privacy of published data, we propose a heuristic quad-tree partitioning method as well as a corresponding privacy budget allocation strategy. Experiments and analysis prove that the adaptive sampling mechanism proposed in this paper can effectively track the trend of dynamic changes of data, and the designed differential privacy method can improve the accuracy of counting query and enhance the availability of published data under the premise of certain privacy intensity. The proposed methods can also be readily extended to other areas of big data release applications.

INDEX TERMS Big location data, privacy preserving data publishing, adaptive sampling, differential privacy, heuristic quad-tree partitioning.

I. INTRODUCTION

Location information is widely used in big data applications such as Mobile Internet, Internet of Vehicles, Intelligent Transportation System, Mobile Social Networks, and Location-based Services. Location information collected and released in real time can help the public understand current traffic conditions, realize communication and sharing based on social networks, perform e-commerce-based diet, play, shopping, and check news, make friends, post information and etc. The immeasurable values brought by the analysis, mining and application of location information has attracted the attention of governments, industries and research departments all over the world.

The existing big location data systems mainly adopt a fixed publishing time interval. For example, in an intelligent transportation system, a GPS device equipped on vehicle reports the location information to the traffic management center every 10-20 seconds. Vehicle detector stations usually report the nearby traffic flows every few minutes. The publishing platform provides statistic location information for users to keep abreast of traffic conditions, to plan personal travel plan, and to get location-based services, etc. Sampling and publishing big location data with a fixed time intervals makes the implementation the easiest and most convenient. However, from the perspective of the availability and effect of published data, it is far from satisfying the user's demand for real-time application of location big data. Figure 1 shows the real-time traffic flow and sampling results with a fixed time interval in a city for one day. It is not difficult to find that although the published data maintains the trend of the original data as a whole, when the time interval Δt is too large, the sampling result is easy to lose the peak and valley values of the traffic flow, or the overall trend has shifted in time (the

The associate editor coordinating the review of this manuscript and approving it for publication was Mahmoud Barhamgi.

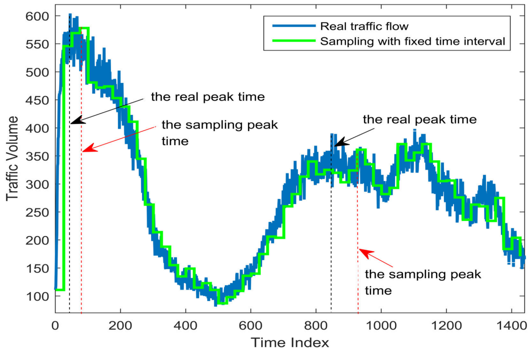


FIGURE 1. Traffic flow and fixed sampling interval.

peak time of the sample result is always earlier or later than the true peak time). It will severely compromise the application performance of location-based big data. For example, users will not be able to schedule their travel at reasonable times. On the contrary, when the time interval Δt is too small, the amount of storage, operations, and calculations required for the publishing of data will increase sharply, resulting in unnecessary waste of resources. Therefore, the reasonable releasing time should be adaptively adjusted according to the change of data volume, so that the sampled value at the releasing time can accurately reflect the dynamic change of location data, and at the same time can balance the amount of system calculation and the availability of published data.

Big location data are closely related to user's private information. By collecting, mining, and analyzing position or trajectory information (time series of position), attackers can not only obtain the location where a user always stays, but also further predict the current position and future trajectory of the user, resulting in the disclosure of personal private information, such as home address, lifestyle, health status, personal interests and income levels. In some serious cases, the leakage of family location may leads to the theft of property, and the leakage of user's trajectory may leads to kidnapping [1], [2].

The privacy protection task for location data can be achieved by cutting off association between users and specific location points or track by the means of anonymity, suppression, perturbations, and encryption [3]–[7]. However, suppressing sensitive location points or encrypting location information will block the application of location big data. Because location-based services need to give the query result or recommendations according to a user's location. Most of the methods based on k-anonymity model and its improved strategy use generalization operations, which have a great impact on the availability of published data. Perturbation methods can achieve privacy protection by adding noise to the real records, and maintain the availability of location information. As one of the research hot-spots of privacy protection perturbation methods, the differential privacy model [8], [9] has been widely used in the privacy protection process of statistical big data. By adding random noise to the original data, the differential privacy model ensures that attackers cannot recognize whether a record exists in the data set. Differential privacy model allows attackers to have infinite computing power and any useful background knowledge, without the need to care about specific attack strategies. In the worst case, even if the attacker obtains all sensitive data except one record, differential privacy model can still guarantee that the attacker cannot judge whether the record is in the dataset.

The differential privacy protection model has a natural match with the application of big location data publishing. The reason is that the large-scaled location data makes the impact of adding or deleting a certain location point on the overall data set very small. This characteristic just coincides the connotation of the differential privacy definition. However, the dynamic update feature of location data also brings new challenges to the application of the differential privacy model, specifically:

- How to dynamically adjust the privacy budget allocation according to the changing location big data, and then maintain a certain privacy protection intensity in each release result, so as to prevent attackers from obtaining user's private information through joint analysis of the published data at different times?
- How to adjust the embedding strategy of random noise according to different distribution of location big data, so as to improve the availability of published data?

In this paper, we discuss the dynamic publishing mechanism and differential privacy protection method for big location data. In a nutshell, the main contributions of this paper are as the following:

- In order to make the published information more accurately reflect the dynamic changing of real big location data, we propose an adaptive sampling mechanism based on proportional-integral-derivation (PID) controller. It can adaptively adjust the length of publishing time intervals according to the amount of updated location information, which not only overcomes the disadvantages of the uniform time interval publishing method, but also improves the availability of published data.
- Aiming at the privacy protection problem of big location data publishing, we propose a heuristic quad-tree partition algorithm based on regional uniformity, and design the corresponding privacy budget allocation strategy to solve the difficulty of determining the stopping condition for top-down space decomposition. It can balance the influence of noise error and non-uniformity error on the query accuracy of published data.
- We conduct extensive experimental studies to demonstrate i) the effectiveness of our proposed adaptive sampling mechanism in tracking the trend of dynamic big location data and ii) the improvement of the differential privacy method in accuracy of the counting queries and availability of released big location data.

The rest of the paper is organized as follows. Section II introduces the preliminaries. Section III provides the basic principle and implementation method of the proposed

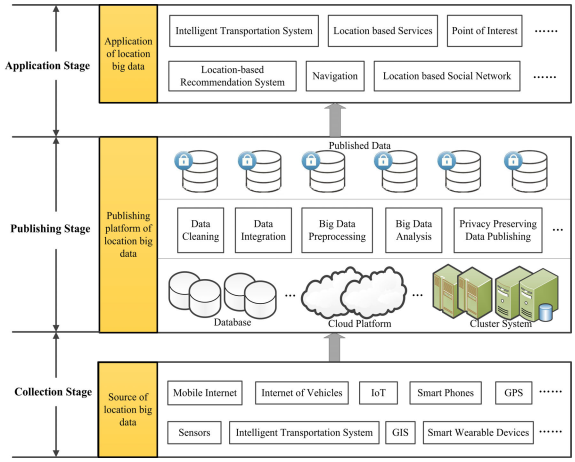


FIGURE 2. **The framework of centralized release of big location data.**

adaptive sampling mechanism. Section IV details the heuristic partition and differential privacy protection algorithm after adaptive sampling. Section V reports a set of empirical studies and the results. Section VI reviews the related works, and Section VII concludes the paper.

II. PRELIMINARIES

A. THE FRAMEWORK

Figure 2 shows the framework of centralized release and privacy preserving of big location data. Massive sensors and intelligent terminals continuously collect location information from various users. After the aggregation, integration, analysis and processing, location data are provided to different fields of scientific research, decision support and public services, which fully reflect the value of location data in various applications such as intelligent transportation systems, location-based services, location-based advertising. Because of the close relationship between location information and personal privacy, releasing location data without privacy protection is very likely to cause user privacy leakage, leading to serious threats on personal reputation, interests and security. The goal of privacy preserving-based location data publishing is to improve the availability of data and achieve various location-based applications while ensuring that user privacy is not compromised.

The data producers can be different kinds of terminals, devices, sensors, vehicles, users, etc. from diverse channels such as the Mobile Internet, Internet of Vehicles, Internet of Things, Sensor Networks and Social Networks. Location data are constantly sent to the publishing platform in different forms, structures and quality. Most of this process uses an honest model, assuming that all the data is sent to a reliable data publisher. In order to protect the privacy information of data producer, the identifier attribute of the original data that can uniquely identify the user is generally deleted, so that the original data is anonymized.

The publishing stage is the bridge connecting the collection stage and the application stage. After cleaning, pre-processing, fusion, analysis and privacy protection processing, location data are fed to various applications according to different needs and methods. The process of releasing location data can be either interactive query methods or non-interactive publishing. The published data can be organized in the form of location information database, statistical reports, news reports, research reports, etc. The publishing stage of big location data adopts a non-honest model because there may be malicious users (called attackers). By collecting large-scale location data, using advanced big data analysis and mining tools, and leveraging background knowledge gained from other sources, attackers could identify specific users from the published data or gain further access to the user's private information.

In addition to the diversity and openness of big data publishing methods, the continuous development of data mining and analysis technology has helped people discover new knowledge and also led to the disclosure of privacy. For the publishers of location data, it is impossible to fully understand the background knowledge of the data recipients, nor to control how the published data will be used. Therefore, the privacy preserving technology used in the data release process is particularly important. Reasonable and effective privacy protection algorithm can improve the accuracy of query and analysis on the published data under the premise of ensuring user privacy, as well as achieve a balance of data privacy and availability.

B. DIFFERENTIAL PRIVACY

Definition 1 (ϵ-Differential Privacy [8], [9]): For any pair of neighboring datasets D and D' (differing on at most one element, written as $\|D - D'\| = 1$) and all $U \subseteq Range(M)$, a randomized algorithm M gives ϵ-differential privacy if:

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

Differential privacy gives a strong guarantee that the presence or absence of an individual will not significantly affect the final output of a query. Parameter ϵ is called the privacy preserving budget. The smaller the value of ϵ, the higher the privacy protection of user information.

Definition 2 (Global Sensitivity): For any pair of neighboring data sets D and D', and a query Q, the global sensitivity of Q is the maximum value between the query results of D and D':

$$GS\left(Q\right) = \max_{D,D'} \|Q\left(D\right) - Q\left(D'\right)\|_1 \tag{2}$$

where $\|Q\left(D\right) - Q\left(D'\right)\|_1$ denotes the first-order norm distance of $Q\left(D\right)$ and $Q\left(D'\right)$.

Definition 3 (Laplace Mechanism): For any algorithm M, if the output result satisfies Formula (3), algorithm M is said to provide ϵ-differential privacy:

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

where $M\left(D'\right)$ is the noisy count of the algorithm on the published data set, and $M\left(D\right)$ is the result on original data set.

Theorem 1 (Serial Composition): For a set of randomized algorithm $\{M_1, M_2, \ldots, M_n\}$, each of M_i $(1 \leq i \leq n)$ satisfies ϵ_i differential privacy on the dataset D, then the composite algorithm of $\{M_1, M_2, \ldots, M_n\}$ can realize $\sum_{i=1}^{n} \epsilon_i$ differential privacy on the same dataset D.

Theorem 2 (Parallel Composition): Suppose the dataset D can be divided into a series of disjoint subsets $\{D_1, D_2, \ldots, D_n\}$ and there is a set of randomized algorithm $\{M_1, M_2, \ldots, M_n\}$ acting on the above subsets, where M_i $(1 \leq i \leq n)$ satisfies ϵ_i differential privacy on the subset D_i. Then the set of randomized algorithm can achieve $max\ \{\epsilon_i\}$ differential privacy on the dataset D.

Definition 4 (Noise Error): In order to prevent attackers from inferring user's specific information through a large number of queries, Laplace 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 noise error:

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

where $C\ (P)$ and $C\ \left(P'\right)$ indicate the original statistical value and the noisy statistical value of a certain area P_i, respectively.

Definition 5 (Non-Uniformity Error): For a query Q, $P_i\ (i = 1, 2, \ldots, m)$ represents all the cells that are intersect with the query rectangle or contained within it, $\underset{1\leqslant i,j\leqslant m\wedge i\neq j}{P_i \cap P_j} = \varnothing$. $r_i\ (i = 1, 2, \ldots, m)$ is the area ratio of the query area to a cell. The non-uniformity error for the query Q can be described as:

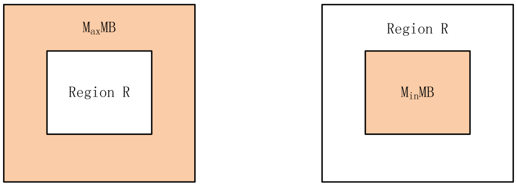
$$NonUni\_err\ (Q) = \left|\sum_{i=1}^{m} r_i \cdot C\ (P_i) - C\ (Q)\right| \tag{5}$$

III. ADAPTIVE SAMPLING METHOD FOR DYNAMIC BIG LOCATION DATA

Big location data mainly come from the GPS positioning information, vehicle data captured by the road induction coil and the cameras, navigation and positioning information of mobile user, etc. For such large-scale, fast-changing, complex and sparse data, if the publishing (i.e., releasing) time interval is fixed, it will face the following dilemma: if the data change faster and the sampling frequency is lower, the sampled data can not reflect the real-time dynamic change of location data, and the quality of the query result returned to user is poor. On the contrary, if the data change slowly while the sampling frequency is high, more computing resources will be wasted, and the corresponding privacy protection process may introduce excessive noise and affect service quality. Therefore, it is necessary to design an adaptive sampling method, which can capture the trend of data updating and reduce the query error of the published data by using the temporal and spatial correlation between sampled datasets.

A. MOVING BOUNDARY

Combined with the distribution characteristics of infrastructures and the behavioral characteristics of the population, location information is always densely distributed within a certain time and space. As a result, between the adjacent releasing times, the distribution features of location data are spatially redundant. The user's moving boundary therefore can be used as a constraint. Assuming that the speed of a mobile user is V, then the extended area of the adjacent time is $X = V \times (t_{i+1} - t_i)$. The moving boundary is defined as follows:



(a) $M_{ax}MB$ (b) $M_{in}MB$

FIGURE 3. Examples of moving boundaries.

Definition 6 (Maximum Moving Boundary $M_{ax}MB$): R represents an area covered by location information on the current time. When location data is updated, all the points located in the extended annular area may enter the area R to form a larger area $M_{ax}MB$, as shown in Figure 3(a).

Definition 7 (Minimum Moving Boundary $M_{in}MB$): R represents the location area of the current time. When the data is updated, some points located in the annular area may disappear, causing the area R shrink into a smaller area $M_{in}MB$, as shown in Figure 3(b).

Generally speaking, the maximum and minimum moving boundaries can be calculated by sampling time intervals and user's moving speed. However, when the latitude and longitude are different, the distance obtained according to user's moving speed and time has different meanings. For example, in the case of equal longitude, the distance is about 1,113 meters per 0.01 degree; while in the case of equal latitude, the distance is about 1,000 meters per 0.01 degree.

B. ADAPTIVE SAMPLING MECHANISM BASED ON PID CONTROLLER

PID controller is the most common form of feedback control, and is mainly used to measure the variation of sampling performance over time [10]. In this section, we design an adaptive sampling PID control method for the dynamic publishing requirements of location data based on moving boundaries. At the initial moment, location data are sampled according to a fixed time interval and obtain a snapshot dataset. The predicted data are then obtained by predicting the maximum and minimum moving boundary according to Definitions 6 and 7. Finally, the feedback error is obtained by comparing the count values of the two datasets, and the PID controller is adjusted to obtain a new sampling time interval according to the feedback error. The principle flow chart of the adaptive sampling control is shown in Figure 4.

After we get the maximum and minimum moving boundaries for the snapshot dataset, the feedback error between the snapshot dataset and the forecast dataset can be obtained

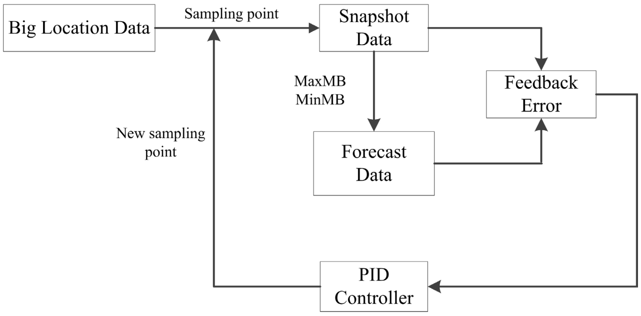


FIGURE 4. Adaptive sampling scheme based on PID.

by comparing the statistic values under the same range. The feedback error is defined as follows:

Definition 8 (Feedback Error): $M_{in}MBCount$ and $M_{ax}MBCount$ represent the count value within the predicted maximum and minimum moving boundaries respectively, *realCount* indicates the actual count value of the region after sampling. The feedback error *FE* can be calculated by using:

(a): if $realCount \leq M_{in}MBCount$,

$$FE = \frac{M_{in}MBCount - realCount}{realCount} \tag{6}$$

(b): if $realCount \geq M_{ax}MBCount$,

$$FE = \frac{realCount - M_{ax}MBCount}{realCount} \tag{7}$$

(c): if $M_{in}MBCount < realCount < M_{ax}MBCount$,

$$meanCount = \frac{M_{in}MBCount + M_{ax}MBCount}{2} \tag{8}$$

$$FE = \frac{|realCount - meanCount|}{realCount} \tag{9}$$

The variation process of the feedback error can reflect the dynamic change of actual location data sampling and publishing. Therefore, the corresponding feedback controller can be designed according to the change of feedback error, which is used to adjust the time for big location data sampling and publishing. By combining the time and spatial correlations of big location data, we redefine the proportional error, integral error and differential error of PID as follows:

Definition 9 (Proportional Error Φ_P): Proportional Error is used to keep the output of the controller proportional to the current error, so that the system deviations can be controlled within a certain range. Once the deviation occurs, the regulator immediately produces a control to reduce the deviation. Let D_P represent the proportional gain, *FE* is the feedback error, the proportional error can be specifically defined as:

$$\Phi_P = D_P \times FE \tag{10}$$

Definition 10 (Integral Error Φ_I): Integral error is used to eliminate the offset by controlling the output rate of the proportional, improving the system's indifference, specifically defined as:

$$\Phi_I = \frac{D_I}{T_I} \times \sum_{j=n-T_I+1}^{n} FE_j \tag{11}$$

where D_I represents the integral gain, T_I represents the number of regions that have errors, and n represents the total number of divided regions.

Definition 11 (Derivative Error Φ_D): Derivative error can reflect the trend or rate of the change of deviation signal, and can speed up the system and reduce the adjustment time by preventing the occurrence of large errors according to the change ratio before the value of deviation signal becomes too large, specifically defined as:

$$\Phi_D = D_D \times \frac{FE_{m_j} - FE_{m_{j-1}}}{m_j - m_{j-1}} \tag{12}$$

where D_D is the derivative gain of the derivative error and m_j $(j = 1, 2, \ldots, n)$ represents a partitioned region.

Definition 12 (PID Error Φ): PID error integrates all the factors from proportional error, integral error, and derivative error, specifically defined as:

$$\Phi = D_P \times FE + \frac{D_I}{T_I} \times \sum_{j=n-T_I+1}^{n} FE_j + D_D \times \frac{FE_{m_j} - FE_{m_{j-1}}}{m_j - m_{j-1}} \tag{13}$$

where $D_P, D_I, D_D > 0$, $D_P + D_I + D_D = 1$.

According to the total error of PID, the new sampling interval can be defined as:

$$T_{new} = max\left\{T_0, T + \alpha\left(1 - e^{\frac{\Phi - \eta}{\eta}}\right)\right\} \tag{14}$$

where T is the original fixed sampling interval, α and η are the specified parameters, α determines the magnitude of change, η is the PID error within the maximum allowable range, T_0 is the given minimum interval.

Algorithm 1 presents the details of the adaptive sampling method based on the PID controller.

IV. PARTITION AND PRIVACY PROTECTION OF BIG LOCATION DATA

Partition is an effective way to release statistical location information, which can provide the query service for the number of users within a certain geographical range, and understand the traffic flow condition within certain areas. Two-dimensional partition method usually divides the space according to a certain index structure. Each index area is represented by the statistical value of the data within the area, which can reduce the disclosure risk of the real location information. By adding differential privacy noise to the statistical value of an index area, the effect of location privacy protection can be further improved.

A. HEURISTIC QUAD-TREE PARTITION AND PRIVACY PROTECTION METHOD

When partitioning the two-dimensional space from top to bottom, it is important to determine the stop condition. If a region has very few points, over-partitioning will create a set of cells with close to zero data points, thus introducing too much noise error. On the other hand, if a region is very

Algorithm 1 Adaptive Sampling Based on PID

```
Require: incoming dataset $D$, moving speed $V$, the original
    fixed sampling interval $T$, minimum sampling interval $T_0$
Ensure: snapshot dataset $SD$, new sampling interval $T_{new}$
$t_1 = 0, t_2 = T$
while the amount of new incoming data point $\neq 0$ do
    snapshot dataset $SD \leftarrow$ select data points from $D$
    within time interval $[t1, t2]$
    partition $SD$ to get subareas $R_i$ $(1 \le i \le n)$
    if $i$ in $n$ then
        get the $realCount$ of region $R_i$
        get $M_{ax}MBCount$ and $M_{in}MBCount$ according to
        moving speed $V$
    end if
    get $FE$ according to Definition 8
    get $\Phi_P$, $\Phi_I$, $\Phi_D$ and $\Phi$ according to Formula (10)-(13)
    get $T_{new}$ according to Formula (14)
    $t_1 \leftarrow t_2$
    $t_2 \leftarrow t_1 + T_{new}$
    return $SD$ and $T_{new}$
end while
```

dense, under-partitioning will increase the non-uniformity error, and reduce the accuracy of count query. Most of the existing partition methods get the proper level of partition (such as the depth of the tree structure) according to some experimental results. Assigning the level of partition based on posteriori results will limit the accuracy of the allocation of privacy budget, and cannot meet the requirement of dynamic publishing of location data. In order to solve this problem, we define the uniformity of a region and design a heuristic quad-tree partitioning algorithm based on region uniformity.

Definition 13 (Regional Uniformity): For a given region S, $den\,(D_1)$, $den\,(D_2)$,…,$den\,(D_i)$ represent the density of sub-areas after multiple direction partition (for example: vertical, horizontal, diagonal, center and periphery). A row vector $V = \{den\,(D_1), den\,(D_2), \ldots, den\,(D_i)\}$ can be constructed, and the uniformity can be expressed as $U = \log_{10}(Var\,(V))$. For the given threshold θ, if Formula (15) is satisfied, the area is uniform.

$$\left| U - log_{10}\left(\frac{den\,(S)}{i}\right)^2 \right| \le \theta \tag{15}$$

Algorithm 2 describes the details of the heuristic quad-tree partition method based on region uniformity. Firstly, we judge the uniformity of the area according to Definition 13. If it is not satisfied, the area will be partitioned into four sub-areas of the same size, and each of the subarea will be traversed in turn according to the depth-first rule and carry out the same uniformity judgement and partition steps, until the distribution characteristics of the current region meet the uniformity condition or the current region area is smaller than the given minimum value.

Algorithm 2 Heuristic Quad-Tree Partition (HQP)

```
Require: snapshot dataset $SD$, minimum region value $M_{in}$
Ensure: special structure $anc$ after heuristic quad-tree parti-
    tion, depth of quad-tree $h$
root $= SD$
$h = 0$
$anc \leftarrow SD$
PreNode $= SD$
if range$(SD) < M_{in}$ then
    return $anc$ and $h$
else
    if $SD$ satisfies the uniformity condition then
        $anc \leftarrow$ mark PreNode as a leave node
        if PreNode is the root then
            return $anc$ and $h$
        else
            set four sub-areas as PreNode one by one
            $anc \leftarrow$ recursive call algorithm HQP
        end if
    else
        partition PreNode with Quad-tree method
        $h = h + 1$
        set four sub-areas as PreNode one by one
        if PreNode is the root then
            return $anc$ and $h$
        else
            $anc \leftarrow$ recursive call algorithm HQP
        end if
    end if
end if
```

B. ALLOCATION AND ADJUSTMENT OF PRIVACY BUDGET

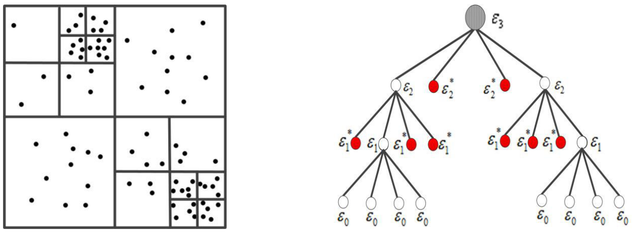
Traditional quad-tree partition method uses the geometric privacy budget allocation strategy [4], that is, the privacy budget increases by $2^{\frac{1}{3}}$ step by step starting from the root node of quad-tree, and the query precision of the leaf nodes is optimal. In order to achieve better query precision of the dynamic release of location data, we improve the traditional geometric budget allocation scheme and apply it into the heuristic quad-tree partition and privacy protection process.

Suppose the total privacy budget is ϵ, the depth of quad-tree obtained by heuristic quad-tree partition algorithm is h. In order to maintain each path from the leaf node to the root node that satisfies ϵ-differential privacy, the geometric budget allocation can be adjusted as follows:

1) For areas that do not satisfy the uniformity condition (which means the nodes need to be partitioned by standard quad-tree structure), the traditional geometric budget allocation method is used:

$$\epsilon_i = 2^{\frac{(h-i)}{3}} \times \epsilon \times \frac{\sqrt[3]{2} - 1}{2^{\frac{(h+1)}{3}} - 1} \tag{16}$$

$$\sum_{i=0}^{h} \epsilon_i = \epsilon. \tag{17}$$



(a) Spatial distribution of heuristic quad-tree partition

(b) Corresponding quad-tree structure and privacy budget allocation

FIGURE 5. Example of heuristic quad-tree partition and privacy budget allocation.

2) For areas that the uniformity condition is satisfied (which means no partitioning is need), the privacy budget can be set to:

$$\epsilon_j{}^* = \sum_{i=0}^{j} \epsilon_i \tag{18}$$

Figure 5 is an example of heuristic quad-tree partition and privacy budget allocation. The red node represents the area where the uniformity condition is satisfied and no further partitioning is needed. According to the adjusted geometric privacy budget allocation method, $\epsilon_1{}^* = \sum_{i=0}^{1} \epsilon_i = \epsilon_0 + \epsilon_1$, $\epsilon_2{}^* = \sum_{i=0}^{2} \epsilon_i = \epsilon_0 + \epsilon_1 + \epsilon_2$. Each of the path from the red leaf node to the root node can achieve $\epsilon_0 + \epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon$ differential privacy, thus ensuring the correctness and stability of privacy budget allocation for the heuristic quad-tree partition algorithm.

In order to further enhance the query accuracy of the released data, the post-processing method [11] is used to adjust the consistency of the noisy count after heuristic quad-tree partition and differential privacy protection. The above objectives can be achieved by two steps:

- Step 1: calculating the estimate value $Z(v)$ for each node v in the heuristic quad-tree structure from bottom to top. For leaf node v, set $Z(v) = C(v')$, otherwise,

$$Z(v) = \frac{m^l - m^{l-1}}{m^l - 1} C(v') + \frac{m^{l-1} - 1}{m^l - 1} \sum_{u \in succ(v)} Z(v) \tag{19}$$

where $C(v')$ represents the noisy count of node v, m represents the number of leaf nodes of the current subtree ($m = 4$ for a quad-tree partition), l is the level at which node v is located (from the leaf node to the upper node $l = 1, 2, \ldots, h+1$), $succ(v)$ is the set of child nodes of node v.

- Step 2: Obtaining the consistency estimate value $H(v)$ for all the nodes from top to bottom. For the root node v, set $H(v) = Z(v)$, otherwise,

$$H(v) = Z(v) + \frac{1}{m}\left[H(u) - \sum_{w \in succ(u)} Z(w) \right] \tag{20}$$

where u represents the parent node of v.

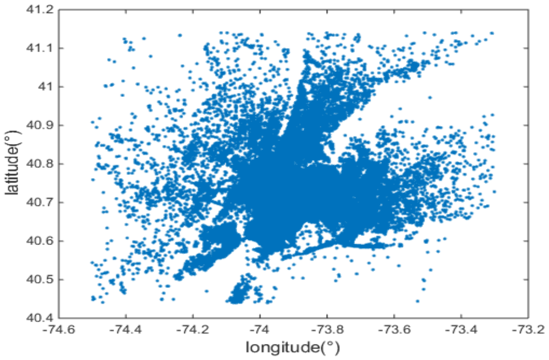


FIGURE 6. Distribution area of the experimental dataset.

C. DYNAMIC DIFFERENTIAL PRIVACY RELEASE ALGORITHM FOR BIG LOCATION DATA

Combined with the above-designed adaptive sampling mechanism and differential privacy partition method, we design the overall process of dynamic publishing and privacy protection algorithm for location data, as shown in Algorithm 3. For the location data that arrive continuously and change constantly, first, we adaptively select the publishing time interval by Algorithm 1, so as to obtain a snapshot of sampled data. Next, we carry out the heuristic quad-tree partition on the snapshot dataset according to Algorithm 2, to get the spatial partition structure and depth of quad-tree which satisfies the uniformity condition. Then, the privacy budget is assigned to nodes of different levels according to the scheme designed by Formula (16)-(18), and the Laplace differential privacy noise is added into the original statistic results. Finally, we post-process the noisy count according to the consistency condition given in Formula (19)-(20) and get the final released data.

Algorithm 3 Dynamic Differential Privacy Release Method

```
Require: incoming dataset $D$, privacy budget $\epsilon$
Ensure: published data
while the amount of new incoming data point $\neq 0$ do
    carry out Algorithm 1 to get snapshot dataset $SD$ and new sampling interval $T_{new}$
    carry out Algorithm 2 on $SD$ to get special structure $anc$ and the depth of quad-tree $h$
    calculate privacy budget $\epsilon_l$ $(1 \le l \le h)$ according to Formula (16)-(18)
    for each node within $anc$ do
        for each level of quad-tree $l$ do
            get the $realCount$ of node region
            noiseCount $= realCount + Laplace\left(\frac{GS}{\epsilon_i}\right)$
        end for
    end for
    published data $\leftarrow$ post-process the noiseCount according to Formula (19)-(20)
end while
```

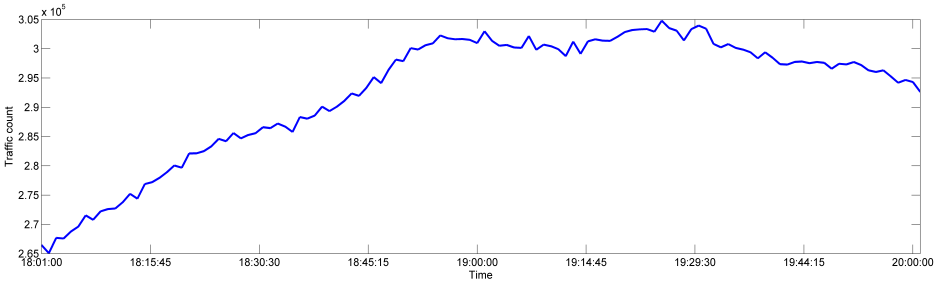


FIGURE 7. Statistics of the experimental datasets.

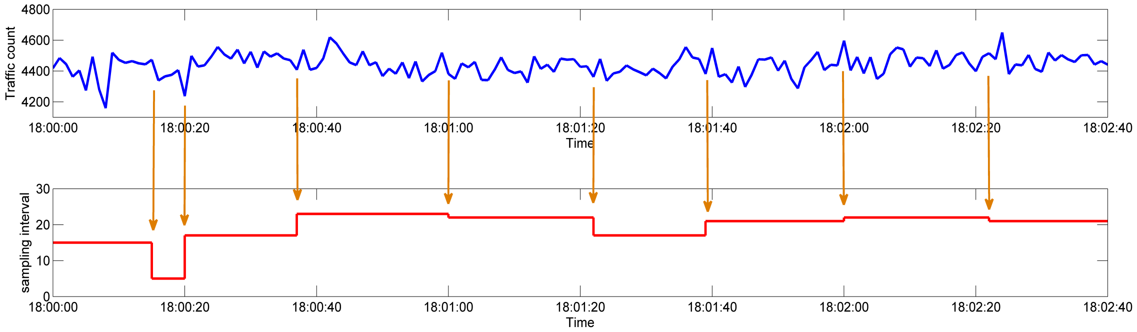


FIGURE 8. Comparison of the original data sequence with adaptive sampling interval.

V. EXPERIMENT AND ANALYSIS

We evaluate the performance of our method through an extensive set of experiments and analysis. Some typical partition algorithms such as UG [2], AG [2], Quad-post [1] and Quad-heu [12] are used to compare with the proposed heuristic quad-tree partition and privacy preserving method.

A. EXPERIMENTAL SETUP

We use the TLC travel record dataset of New York City[1] in our experiments. The experimental datasets select the latitude and longitude information between the period from 18:00 to 20:00 each day in 2015. Distribution area of the dataset is shown in Figure 6. The experimental platform is Alibaba Cloud Server ECS (ecs.r5.xlarge: 4-core CPU, 32GB RAM, 100G SSD cloud disk, Windows Server 2008 R2 Enterprise Edition image). All the algorithms are programmed by MATLAB R2015b. The location information is arranged in chronological order to obtain the statistical graph in the statistical sense, as shown in Figure 7. During this period, the statistical information reflects a certain trend, representing the lifestyle of the public during the corresponding time period.

B. EXPERIMENTAL RESULTS

1) ADAPTIVE SAMPLING EFFECT OF REAL LOCATION BIG DATA

In the experiment of adaptive sampling mechanism based on PID controller, we set V=17m/s(equal to 60 km/h), D_P=0.9, D_I=0.1, D_D=0, η=0.2, α=15, T=15. In order to prevent computation and storage resource consumption caused by small sampling intervals, we set T_0=5. Figure 8 compares the results of real statistical information of the data with adaptive sampling during the time range from 18:00:00 to 18:02:40. It can be seen from the figure that the sampling interval obtained by the proposed adaptive sampling mechanism can capture the trend of data changes. When the statistical data fluctuates significantly, the sampling interval is reduced and the the sampling rate is increased. When the amount of data change is relatively flat, the sampling interval increases, the sampling rate decreases, and unnecessary calculation and storage steps are avoided. At the beginning of the experiment (from 18:00:00 to 18:00:15), the first data snapshot is obtained by a fixed sampling interval (T=15), so the released data do not reflect the violent changes during this period. This also verifies the limitations of the evenly spaced sampling and publishing method.

Figures 9 and 10 show the results of sampling with fixed time interval and adaptive sampling over the same time period. The experiment period is 2 minutes long and the mean value change is not obvious. The sampling method with fixed time interval obtains 8 snapshots, and the total amount of each snapshot is basically the same. However, the adaptive sampling method gets 7 snapshots, the total amount of the snapshot changes significantly, and the trend is closer to the original data. The experimental results further prove that the adaptive sampling method has better ability to capture the details of the dynamic changing location data.

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

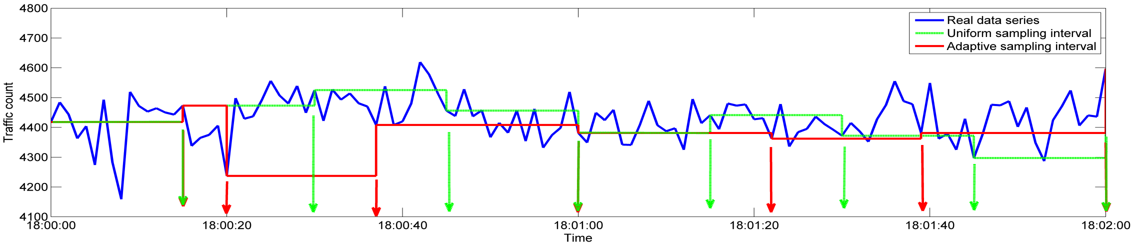
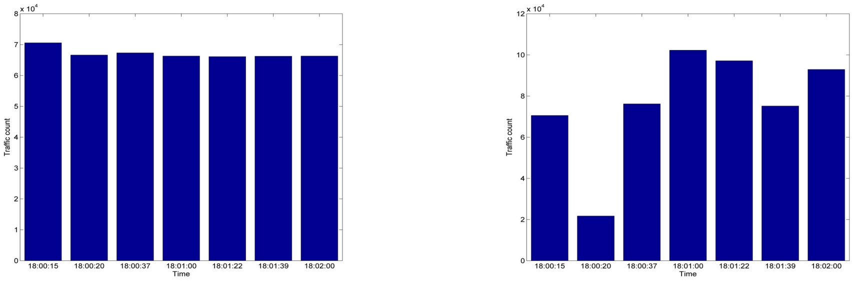


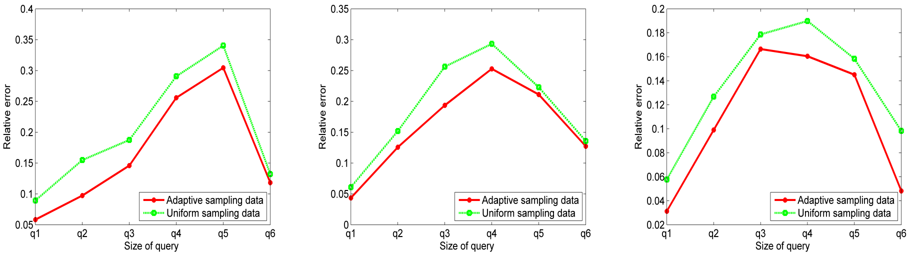
FIGURE 9. Comparison of the original data sequence with adaptive sampling interval.



(a) Uniformly sampled data set

(b) Adaptive sampled dataset

FIGURE 10. Comparison of the number and trends of snapshots between uniform sampling and adaptive sampling.



(a) $\epsilon = 0.1$

(b) $\epsilon = 0.5$

(c) $\epsilon = 1.0$

FIGURE 11. Comparison of query errors between different sampling methods.

2) COMPARISON OF QUERY ERRORS BETWEEN DIFFERENT SAMPLING METHODS

Next, we carry out the partition step on the snapshots at each sampling time and add the differential privacy noise based on Laplace mechanism to achieve privacy protection for the published data. Relative error of different scales is used to represent the effects of the sampling method with fixed time interval and the adaptive sampling method. To be fair, we use the same heuristic quad-tree partition method and differential privacy intensity. The results are shown in Figure 11. During the experiment, the Laplace noise is changed by adjusting the privacy budget ϵ. The calculation method of relative error is defined as follows:

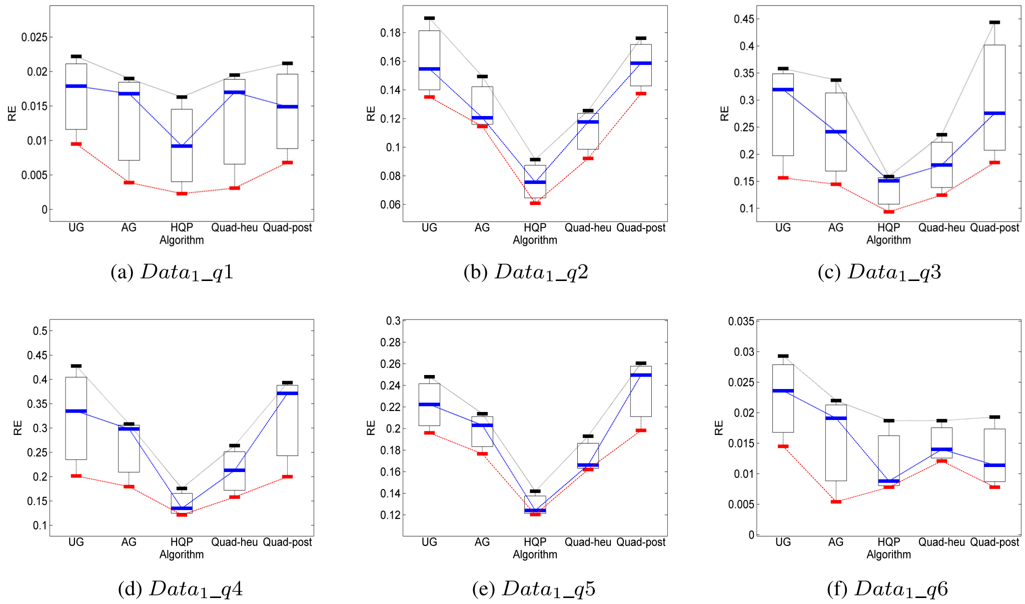
$$RE = \frac{\left|Q\left(D\right) - Q\left(D'\right)\right|}{max\{Q\left(D\right), \rho\}} \tag{21}$$

where $Q\left(D\right)$ is the query result of the real statistical value, $Q\left(D'\right)$ is the statistical value after adding Laplace noise, $\rho = 0.001 \times |D|$, $|D|$ represents the size of the experimental dataset.

As can be seen from Figure 11, under the same privacy budget, the published data of the adaptive sampling mechanism has a smaller relative error at various query sizes. Therefore, it is possible to provide location-based services with higher quality.

3) COMPARISON OF DIFFERENT PARTITION METHODS

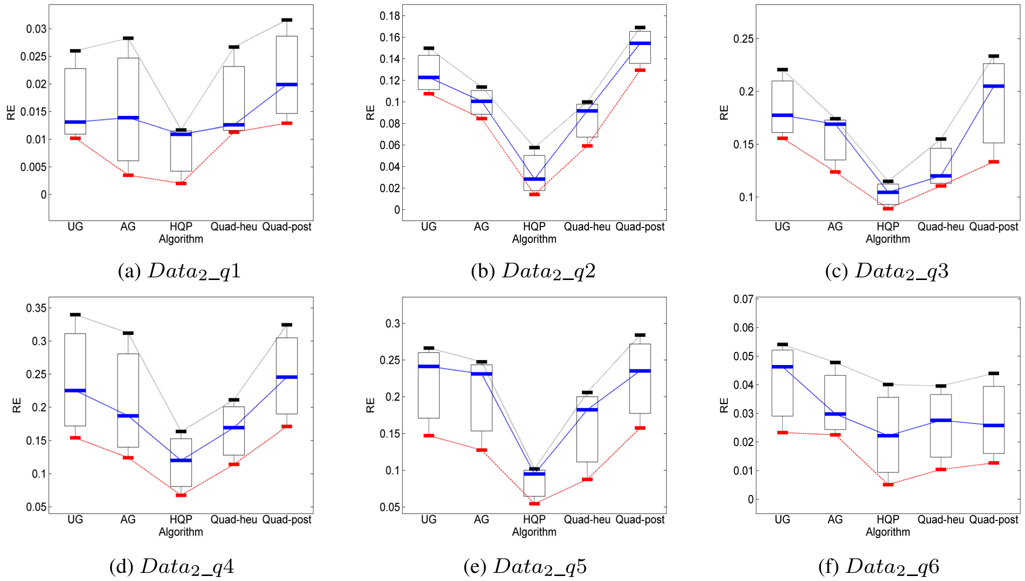
In order to verify the effectiveness of heuristic quad-tree partition and privacy budget allocation method, we compare the proposed algorithm with some typical differential privacy partition algorithms, such as UG, AG, Quad-post and Quad-heu. To be fair, all the partition methods use the same snapshot dataset and differential privacy intensity. During the experiment, Laplace noise was changed by adjusting the privacy budget ϵ. Relative errors of different scales are used to evaluate the performance of different algorithms. The experimental datasets ($Data_1$ and $Data_2$) are selected from snapshot datasets after adaptive sampling. Table 1 shows the overall situation and query range information of the experimental datasets, where $q1 - q6$ represents the coverage of longitude



(a) $Data_1\_q1$ (b) $Data_1\_q2$ (c) $Data_1\_q3$

(d) $Data_1\_q4$ (e) $Data_1\_q5$ (f) $Data_1\_q6$

FIGURE 12. Comparison of query precision on *Data*$_1$.



(a) $Data_2\_q1$ (b) $Data_2\_q2$ (c) $Data_2\_q3$

(d) $Data_2\_q4$ (e) $Data_2\_q5$ (f) $Data_2\_q6$

FIGURE 13. Comparison of query precision on *Data*$_2$.

and latitude. During the experiments, the grid-based partition algorithm UG and AG adopt the partition granularity of $c = 10$, and the tree-based partition algorithm Quad-post and Quad-heu set the depth $h = 6$.

Figures 12 and 13 show the range of relative error of different partition algorithms over different query scales. Among them, the black line, the blue line and the red line respectively indicate the relative error of the privacy budget $\epsilon = 0.1$, $\epsilon = 0.5$, and $\epsilon = 1.0$. As can be seen from the figure, the relative errors of the proposed algorithm are lower than the other algorithms under almost all query scales. For example, on the dataset $Data_1$, when the query range is $q1$ and $\epsilon = 0.1$, the relative error of the proposed algorithm is 0.0163, UG achieves 0.0222, AG achieves 0.0190, Quad-heu achieves 0.0195 and Quad-post achieves 0.0212. The relative error of

TABLE 1. Experimental dataset and query ranges.

Parameter	Data sets	
	Data$_1$	Data$_2$
Total data amount	106561	70428
Coverage	0.3×0.36	
q1	0.00375×0.004375	
q2	0.0075×0.00875	
q3	0.015×0.0175	
q4	0.03×0.035	
q5	0.06×0.07	
q6	0.12×0.14	

the proposed algorithm is reduced by 36%, 17%, 20%, and 30% compared with other algorithms. When the query range is $q5$ and $\epsilon = 1.0$, the relative error of proposed algorithm is 0.1203, the UG algorithm is 0.1960, the AG algorithm is 0.1766, the Quad-heu algorithm is 0.1621, and the Quad-post algorithm is 0.1983. The relative error is reduced by 63%, 47%, 35%, and 65% respectively. The results of Figure 13 also give the similar conclusions.

It can also be observed from Figure 12 and 13 that as the privacy budget increases, the relative errors of different algorithms gradually decrease. Because for the same sensitivity, the increase of privacy budget leads to a reduction of Laplace noise, so the deviation between the published results and the actual data is reduced, and the relative error of the query is also reduced. The above experimental results show that compared with the existing partition and publishing algorithms, the proposed algorithm has smaller relative errors in different datasets, different privacy budgets and different query ranges, which can meet the high-precision query requirements of user while using the location-based big data services.

C. ANALYSIS OF THE PRIVACY PROTECTION INTENSITY

This section discusses the proof that statistical release of big location data using the heuristic quad-tree partition method and the dynamic privacy budget allocation strategy, can provide ϵ differential privacy intensity for any range of query area Q.

Proof: For the counting query within an arbitrary range Q proposed by a user, there are generally the following two situations:

- The query range Q falls completely within a certain node area after heuristic quad-tree partition. Let ϵ_Q indicate the privacy protection intensity of the node area, assuming that the node is located in the layer l of the heuristic quad-tree structure (l=0 represents the leaf node, l=h represents the root node), according to the serial combination property of differential privacy, the privacy intensity of the node is $\epsilon_Q = \sum_{j=0}^{l} \epsilon_j$. If the layer j is not a heuristic quad-tree partitioning area, then the node's privacy budget $\epsilon_j = \epsilon_j^* = \sum_{i=0}^{j} \epsilon_i$ and $\epsilon = \sum_{i=0}^{j} \epsilon_i$. If the layer j has been heuristic partitioned, the node's privacy budget will be $\epsilon_j = 2^{\frac{(h-j)}{3}} \epsilon \frac{\sqrt[3]{2}-1}{2^{\frac{(h+1)}{3}}-1}$, $\epsilon = \sum_{j=0}^{h} \epsilon_j$. Therefore, the overall privacy budget for the query range Q is $\epsilon_Q = \epsilon$.
- The query range Q contains $n\,(2 \leq n \leq N)$ different node regions. Let ϵ_{Q_i} $(i = 1, 2, \ldots, n)$ indicate the privacy protection intensity of different node areas, according to the parallel combination property of differential privacy, the privacy protection intensity of query range Q will be $\epsilon_Q = max\left\{\epsilon_{Q_i}\right\}$. For $\forall Q_i$, the intensity of privacy protection is exactly the same as situation (1), that is $\epsilon_{Q_i} = \epsilon$, so we can also get the result $\epsilon_Q = \epsilon$. □

VI. RELATED WORK

Dynamically changing big data can be seen as a kind of data streams. Sampling and publishing according to certain strategies is the main method to realize the dynamic release of big data. Obviously, this kind of method can save a lot of resources, but presents the research problem of effective and efficient data sampling [13]. Within this area, regional sampling, reservoir sampling, priority sampling, as well as the combination methods [14]–[18] have been proposed. However, there are limitations and difficulties of refining the characteristics of stream data. Some research proposed sliding window sampling methods to solve the transient characteristics of streaming data [19], [20]. Although these sampling methods have obvious advantages in data processing, there is no complete probability structure, which makes hard to analyze the sampling error and accuracy.

Numerous techniques have been proposed to protect the privacy of location information, such as dummy locations, k-anonymity, obfuscation methods, and differential privacy. K-anonymity methods [3], [4] replace the user's exact location point with a spatial region containing the location point, and combine the k-anonymity model to generalize the user coordinates into an area containing at least k users, thereby protecting the user's precise location. The dummy location method [5] allows a user to generate fake location points according to his privacy protection requirement, and sends the fake location points together with the real ones to the service provider, so that the true location of the user cannot be easily guessed by malicious attackers. This kind of method is widely used to protect user location privacy in a single query. For a large number of cellular network mobile data, the DP-Where method [6] was designed to achieve differential privacy protection by adding controlled noise to the uniform grid structure. The Kd-PPDP algorithm [7] uses the square sum error to measure the uniformity of the current mesh after adaptive meshing and noise addition, and merges the adjacent regions to reduce the noise error. Differential privacy model protects location privacy by introducing a trusted third party and partition the two-dimensional space according to certain index structures. Some typical methods such as tree-based partitioning methods [21], [22], grid-based partitioning method [23], and hybrid structure partitioning method [24] are all aimed at reducing the non-uniform errors and noise errors generated during the partition process. Xiong et al. [25] presented a differential private allocation mechanism for reward-based spatial crowdsourcing. They presented a contour plot to characterize location distribution, which firstly partitions the entire area into some disjoint cells, and then connects the cells with the same noisy count to form a larger region. Yang *et al.* [26] proposed a data release mechanism for mobile crowdsensing with differential privacy to provide rigorous protection of worker locations. They designed a recursive partitioning process based on worker density, and applied the non-uniform quad-tree partitioning technique to divide the cells according to the density of workers.

VII. CONCLUSION

Large-scale and variability of big location data present new challenges to privacy protection, which may break the privacy model established during the dynamic publishing process, resulting in the invalidation of privacy protection method, and increasing the difficulty of partitioning and privacy budget allocation. If the published data cannot reflect the trend and magnitude of big data changes, it will lose the real-time performance and affect the quality and availability. In order to better provide users with high-quality location-based services, this paper proposes an adaptive sampling mechanism and privacy protection method for the release of location data. The adaptive sampling method based on the proportional-integral-derivative (PID) controller is designed to determine a reasonable release time for dynamical location data. A heuristic quad-tree partitioning method and a privacy budget allocation strategy are proposed to perform differential privacy protection on the published data. The experiments show that our proposed method can improve the accuracy of counting query and enhance the availability of location-based big data under the premise of certain privacy intensity.

There are some limitations in the proposed method. For example, the data trend on previous interval is not considered when sampling, and only the current snapshot is used to predict the amount of data on the next moment. In our future work, we will focus on the prediction based on the data from adjacent moments. Moreover, the real-time update feature of location data can be regarded as a set of uncertain data streams that grow indefinitely over time. Therefore, privacy protection of streaming location-based big data is another research direction for our future work. This can take into account the real-time nature of data updates, as well as the effectiveness of the privacy model established in the dynamic process.

REFERENCES

[1] G. Cormode, C. Procopiuc, D. Srivastava, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Proc. 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 20–31.

[2] W. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," in *Proc. 29th Int. Conf. Data Eng. (ICDE)*, Apr. 2013, pp. 757–768.

[3] J. Zhang, X. Xiao, and X. Xie, "PrivTree: A differentially private algorithm for hierarchical decompositions," in *Proc. 36th ACM Int. Conf. Manage. Data*, Jun. 2016, pp. 155–170.

[4] A. Ye, Y. C. Li, and X. Li, "A novel location privacy-preserving scheme based on l-queries for continuous LBS," *Comput. Commun.*, vol. 98, pp. 1–10, Jan. 2017.

[5] S. Hayashid, D. Amagata, T. Hara, and X. Xie, "Dummy generation based on user-movement estimation for location privacy protection," *IEEE Access*, vol. 6, pp. 22958–22969, 2018.

[6] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, "DP-WHERE: Differentially private modeling of human mobility," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 580–588.

[7] S. Huang, T. Chen, Q. Lu, Y. Wu, and S. Ye, "Differentially privacy two-dimensional dataset partitioning publication algorithm based on kd-tree," *J. Shandong Univ. (Eng. Sci.)*, vol. 45, no. 1, pp. 24–29, 2015.

[8] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata, Lang., Program.*, 2006, pp. 1–12.

[9] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.

[10] M. King, *Process Control: A Practical Approach*. Chichester, U.K: Wiley, 2011.

[11] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 1021–1032, Sep. 2010.

[12] Y. Wu, Q. Lu, J. Cai, and X. Wang, "Differential privacy two-dimensional data partitioning publication algorithm based on quad-tree," *J. Huazhong Univ. Sci. Technol.(Natural Sci. Ed.)*, vol. 44, no. 3, pp. 99–104, 2016.

[13] Z. Wang and S. Xu, "Research on the method of sampling and storage for a class of data stream," *Statist. Inf. Forum*, vol. 33, no. 10, pp. 16–22, 2018.

[14] M. Greenwald and S. Khanna, "Space-efficient online computation of quantile summaries," *ACM SIGMOD Rec.*, vol. 30, no. 2, pp. 58–66, 2001.

[15] M. Balazinska, H. Balakrishnan, and M. Stonebraker, "Load management and high availability in the Medusa distributed stream processing system," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2004, vol. 16, no. 1, pp. 929–930.

[16] G. Cormode, S. Muthukrishnan, and I. Rozenbaum, "Summarizing and mining inverse distributions on data streams via dynamic inverse sampling," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, vol. 6, no. 1, pp. 25–36.

[17] N. Alon, N. Duffield, C. Lund, and M. Thorup, "Estimating arbitrary subset sums with few probes," in *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, Jun. 2005, vol. 34, no. 1, pp. 317–325.

[18] Y. Wang and S. Wang, "Challenges and opportunities of sampling survey in the age of big data," *Statist. Inf. Forum*, vol. 31, no. 189(06), pp. 33–36, 2016.

[19] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Proc. 21th ACM Sigmod-Sigact-Sigart Symp. Princ. Database Syst.*, 2002, vol. 5, no. 5, pp. 1–16.

[20] D. Tang, C. Liu, Q.-J. Yue, and J.-Y. Zhang, "Adaptive weighted random sampling algorithm based on time sliding window," *J. Dalian Univ. Technol.*, vol. 52, no. 5, pp. 772–775, 2012.

[21] J. Wang, R. Zhu, S. Liu, and Z. Cai, "Node location privacy protection based on differentially private grids in industrial wireless sensor networks," *Sensors*, vol. 18, no. 2, pp. 410–424, 2018.

[22] J. Wang, S. Liu, Y. Li, H. Cao, and M. Liu, "Differentially private spatial decompositions for geospatial point data," *China Commun.*, vol. 13, no. 4, pp. 97–107, 2016.

[23] Q. Li, Y. Li, G. Zeng, and A. Liu, "Differential privacy data publishing method based on cell merging," in *Proc. IEEE 14th Int. Conf. Netw., Sens. Control (ICNSC)*, May 2017, pp. 778–782.

[24] Y. Yan, X. H. Hao, and L. Zhang, "Hierarchical differential privacy hybrid decomposition algorithm for location big data," *Cluster Comput. J. Netw. Softw. Tools Appl.*, 2018, doi: 10.1007/s10586-018-2125-z.

[25] P. Xiong, L. Zhang, and T. Zhu, "Reward-based spatial crowdsourcing with differential privacy preservation," *Enterprise Inf. Syst.*, vol. 11, pp. 1500–1517, Nov. 2016.

[26] M. Yang, T. Zhu, Y. Xiang, and W. Zhou, "Density-based location preservation for mobile crowdsensing with differential privacy," *IEEE Access*, vol. 6, pp. 14779–14789, 2018.

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

LIANXIU ZHANG received the B.Eng. degree from the University of Tarim, in 2017. She is currently pursuing the master's degree with the School of Computer and Communication, Lanzhou University of Technology, China. Her research interests include network and information security, and privacy preservation technology.

QUAN Z. SHENG received the Ph.D. degree in computer science from the University of New South Wales, in 2006. He is a Full Professor and the Head of the Department of Computing, Macquarie University. His research interests include big data analytics, service computing, distributed computing, the Internet computing, the Internet of Things, and Web of things. He was a recipient of the AMiner's Most Influential Scholar Award on IoT, in 2019, the ARC Future Fellowship, in 2014, the Chris Wallace Award for Outstanding Research Contribution, in 2012, and the Microsoft Research Fellowship, in 2003. He is the author of more than 370 publications. He is a member of the ACM.

BINGQIAN WANG received the B.Eng. degree from Lanzhou University of Technology, China, in 2018, where she is currently pursuing the master's degree with the School of Computer and Communication. Her research interests include privacy preservation and dynamic modeling of big data publishing systems.

XIN GAO received the B.Eng. degree from Harbin Normal University, in 2013. She is currently pursuing the master's degree with the School of Computer and Communication, Lanzhou University of Technology, China. Her research interests include information security and dynamic clustering.

YIMING CONG received the B.Eng. degree from the Harbin University of Science and Technology, in 2016. He is currently pursuing the master's degree with the School of Computer and Communication, Lanzhou University of Technology, China. His research interest includes privacy-preserving data publishing.