


<b>ITC 4/47</b> Journal of Information Technology and Control Vol. 47 / No. 4 / 2018 pp. 639-654 DOI 10.5755/j01.itc.47.4.19320	<b>Location Data Record Privacy Protection Based on          Differential Privacy Mechanism</b>	
	Received 2017/10/19	Accepted after revision 2018/11/02
	 <a href="http://dx.doi.org/10.5755/j01.itc.47.4.19320">http://dx.doi.org/10.5755/j01.itc.47.4.19320</a>	

# Location Data Record Privacy Protection Based on Differential Privacy Mechanism

**Ke Gu, Lihao Yang**

Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha 410114, China; School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha 410114, China; School of Information Science and Engineering, Central South University, Changsha 410083, China; e-mails: gk4572@163.com

**Bo Yin**

School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha 410114, China; e-mails: 16574307@qq.com

**Corresponding author:** gk4572@163.com; 16574307@qq.com

Now many location data applications have facilitated people's daily life. However, publishing location data may divulge individual sensitive information. Currently many existing privacy protection schemes cannot provide the balance of utility and protection. Furthermore, as the records about location data may be discrete in database, some existing privacy protection schemes are difficult to protect location data information in data mining. In this paper, our works mainly focus on providing a framework for the privacy protection of location data mining. We propose a location data record privacy protection scheme based on differential privacy mechanism, which employs the structure of multi-level query tree to query and publish location data on database. Our proposed location data privacy protection scheme may discover the relationship of location data from database and protect location data mined. As accessing location preference of user may be related to private (sensitive) location, it is very important to protect highly frequent accessing location data when location data are mined. So, our proposed scheme provides a mechanism to protect highly frequent accessing location data (or accessing location preference of user) by distorting accessing frequencies. In the proposed scheme, we first construct the structure of multi-level query tree from database, then we make double processes of selecting data according to accessing frequencies by the exponential mechanism and one process of adding noises to accessing frequencies by the Laplace's mechanism on the multi-level query tree. Compared with the other schemes, the experiments show the data availability of the proposed scheme is higher and the privacy protection of the scheme is effective.

**KEYWORDS:** location data record, accessing frequency, differential privacy protection, multi-level query tree, Laplace's mechanism.

## 1. Introduction

### 1.1. Background

With the rapid development of computer and network, data mining and data analysis play the increasingly important roles in our social life. The huge amounts of data (such as big data) can bring many application services to our society, such as location data, health data, food data and traffic safety data. Location data is a kind of position information with large scale and rapid change, which mainly comes from vehicle networks, mobile devices and social networks. Now many applications of location data have facilitated people's daily life, thus location data service is called as a kind of new mobile computing service. Currently, it is the key of developing location data services that we must be able to learn and understand position information [32]. However, location data are mainly collected and disseminated by mobile equipment, where many mobile devices and mobile communication technologies must integrate geographical data and individual information into location data. Thus, location data may contain individual privacy information, personal health status, social status and behavior habits. Then mining location data may divulge individual sensitive information so as to influence people's normal life. The works [26, 40] summarized five kinds of commonly used moving object positioning method [1, 13, 16, 20, 33, 34, 37, 41] and three kinds of accessing individual location information approach [12, 17, 29].

Presently, it is the key of location data privacy protection that how to protect sensitive information while providing location service on data mining. Namely, we must find a compromising approach between service and protection. However, many existing privacy protection schemes cannot provide the balance of utility and protection. Furthermore, as the records about location data may be discrete in database<sup>1</sup>, some existing privacy protection schemes are difficult to protect location data in data mining. Therefore, we focus on finding an efficient privacy protection scheme for location data mining in this paper.

<sup>1</sup> In real world, location data may not be discrete. In this paper, our focus is the combination of location data and accessing frequency. Because the combination of location data and accessing frequency is not continuous in database, we consider that the records about location data are discrete.

### 1.2. Our Contributions

In this paper, we propose a location data record privacy protection scheme, which employs the structure of multi-level query tree to query and publish location data. Our proposed location data privacy protection scheme may discover the relationship of location data from database and protect location data mined. As accessing location preference of user may be related to private (sensitive) location, it is very important to protect highly frequent accessing location data when location data are mined. Our proposed scheme provides a mechanism to protect highly frequent accessing location data (related to accessing location preference of user) by distorting accessing frequencies. In the proposed scheme, we first construct the structure of multi-level query tree from database, and then we make double processes of selecting data on accessing frequencies by the exponential mechanism and one process of adding noises to accessing frequencies by the Laplace's mechanism on the multi-level query tree. Additionally, compared with the other related schemes, the experiments show the data availability of the proposed scheme is higher and the privacy protection of the scheme is effective. Our contributions are as follows:

- 1 In our proposed scheme, we construct the structure of multi-level query tree from database. We first use the query tree to represent the result of queried location data, and then we add the noises into the query tree and publish the new query tree as the final result. Our proposed scheme employs the multi-level query tree to show the combination of location data and accessing frequency. Such a method has the following advantages: (a) it can maintain the relationship of location data (as Figure 2) where there is no damage to the original structure of data; (b) it can improve the protection effectiveness of location data: when we add the noises to the corresponding tree nodes, it may protect the relationship of location data and the combination of location data and accessing frequency.
- 2 We make double processes of selecting data according to accessing frequencies by the exponential mechanism and one process of adding noises to accessing frequencies by the Laplace's mechanism on the multi-level query tree. In the double processes of selecting data, the first selection is based on accessing frequency (or support count), where  $n$

location data records whose accessing frequencies are greater than a specified value are selected from the multi-level query tree; the second selection is based on the exponential mechanism, where  $k$  location data records are selected from the  $n$  location data records. In the process of adding noises, noises are added into the accessing frequencies of the  $k$  location data records by the Laplace's mechanism. Such a method minimally distorts the true accessing frequencies of location data records so as to protect some sensitive location data records, where the attackers are difficult to judge the accessing location preference of user.

- 3 The experiments show the running time of the proposed algorithms is less and the privacy protection of the proposed scheme is effective. In addition, by computing true positive rate, false positive rate, accurate rate and false reject rate, the experiments show the data availability of the proposed scheme is higher.

### 1.3. Outline

The rest of this paper is organized as follows. In Section 2, we discuss the related works about privacy protection. In Section 3, we review the related definitions and theorems. In Section 4, we propose an efficient location data record privacy protection scheme, which is based on differential privacy mechanism. In Section 5, we analyze the correctness and security of the proposed scheme. In Section 6, we analyze and show the efficiency of the proposed scheme by the experiments. Finally, we draw our conclusions in Section 7.

---

## 2. Related Work

Currently many privacy protection schemes are being widely used in many fields, such as secure communication, social network, data mining and so on. The works [35, 36] first proposed the  $k$ -anonymity model to protect social network, whose anonymity protection methods mainly include generalization [14, 38], compression, decomposition [47], replacement [50] and interference. Based on the works of [35, 36], many other  $k$ -anonymity protection methods [2, 22, 23, 30, 39, 45, 46, 48] were also proposed. However, the works [2, 21, 49, 52] proved that some anonymous protection methods cannot protect sensitive data very well.

De Cristofaro et al. [8] proposed a privacy-encrypted protection scheme. Although their scheme can ensure data security, data utility is decreased. The existing location data privacy protection methods [4, 32] are mainly classified to three categories: the heuristic privacy-measure methods, the probability-based privacy inference methods and the privacy information retrieval methods. The heuristic privacy-measure methods are mainly to provide the privacy protection measure for some no-high required users, such as  $k$ -anonymity [19],  $t$ -closing [3],  $m$ -invariability [27] and  $l$ -diversity [25]. The information retrieval privacy protection methods may result in no data can be released, and these methods have high overhead. Additionally, the probability-based privacy inference methods can achieve better data utility under certain conditions, but the effectiveness of the methods depends on original data availability. Further, the three kinds of method are based on a unified attack model [32], which depends on certain background knowledge to protect location data. However, with the increase of background knowledge got by the attackers, these methods could not always effectively protect location data. The works [22, 23, 30, 35, 38, 39, 45, 46] showed the shortages of the relationship-privacy protection methods. Wang and Liu [44] analyzed a variety of privacy threat models and tried to optimize the effectiveness of the obtained data while preventing different types of reasoning attack. Gedik and Liu [15] proposed the first effective location-privacy preserving mechanism (LPPM) that enables a designer to find the optimal LPPM for a location-based service. Such a LPPM can maximize the expected distortion (error) when the adversary incurs in reconstructing the actual location of a user. Presently, it is the key of protecting location data to provide a privacy protection method that is not sensitive to background knowledge. Based on the requirement, differential privacy protection technology can exactly satisfy it.

Differential privacy is a kind of strong privacy protection method, which is not sensitive to background knowledge. Li et al. [24] proposed an approach with differential privacy called PrivBasis, which leverages a novel notion of basis set. They introduced the algorithm for privately constructing a basis set and then using it to find the most frequent item-sets. Wang et al. [42] proposed a novel scheme with differential privacy, which directly searches for maximal frequent

item-sets and subsequently adds their sub-item-sets to the results without additional privacy budget consumption. Li et al. [28] proposed a compressive mechanism for differential privacy, which is based on compressed sensing theory. Their mechanism is to consider every data as a single individual, but it undermines the relationship of data so as to be not suitable to protect location data. Ouyang et al. [32] proposed a differential privacy-based transaction data publishing scheme. Their method establishes the relationship of transaction data items by a query tree and adds noises to the query tree based on the compressive mechanism and the Laplace's mechanism. However, it is difficult to measure the effectiveness of their method on privacy protection. Zhang et al. [51] proposed an accurate method for mining top- $k$  frequent data records under differential privacy. In their scheme, the exponential mechanism is used to sample top- $k$  frequent data records, and then the Laplace's mechanism is used to generate noises to distort original data. Although the effectiveness of their method may accurately be measured on privacy protection, their method neglects the relationship of transaction data items.

Based on differential privacy mechanism, many related technologies are used to protect location data. He et al. [18] proposed a synthetic system based on GPS path, which can provide strong differential privacy protection mechanism. The proposed system gets and protects different speed trajectory by using a hierarchical reference method to isolate the original trajectory. Chatzikokolakis et al. [6] proposed a predictive differential-private mechanism for location privacy, which can offer substantial improvements over the independently applied noise. Their work showed that the correlations in the trace can be exploited in terms of a prediction function that tries to guess the new location based on the previously reported locations. In addition, their work tested the quality of the predicted location. Chatzikokolakis et al. [7] also showed a formal notion of privacy that protects the user's exact location—"geo-indistinguishability". In [7], they proposed two mechanisms to protect the privacy of user when dealing with location-based services. They extended their mechanisms to limit the degradation of the privacy guarantees due to the correlation between the points. Bindschaedler and Shokri [5] presented a

synthesizing plausible privacy-preserving location tracing scheme. Wang et al. [43] proposed a real-time spatio-temporal crowd-sourced data publishing scheme with differential privacy.

## 3. Preliminaries

### 3.1. Differential Privacy

Differential privacy protection achieves privacy protection target by making data distortion, where the common approach is to add noises into querying result. The purpose of differential privacy protection is to minimize privacy leakage and to maximize data utility [9, 11]. Currently differential privacy protection has two main methods [10, 31]—the Laplace's mechanism and the exponential mechanism.

*Laplace's mechanism:* Dwork et al. [10] proposed a protection method for the sensitivity of private data, which is based on the Laplace's mechanism. Their method distorts the sensitive data by adding the Laplace's distribution noises to the original data. Their method may be described as follows: the algorithm  $M$  is the privacy protection algorithm based on the Laplace's mechanism, the set  $S$  is the noise set of the algorithm  $M$ , and the input parameters are the data set  $D$ , the function  $Q$ , the function sensitivity  $\Delta Q$  and the privacy parameter  $\epsilon$ , where the set  $S$  approximately subjects to the Laplace's distribution ( $\frac{\Delta Q}{\epsilon}$ ) and the mean (zero), as shown in the formula (1):

$$\Pr[M(Q,D)=S] \propto \exp\left(\frac{\epsilon}{\Delta Q} \times |S-Q(D)|\right). \quad (1)$$

In their method, the probability density function of added function of noises subjecting to the Laplace's distribution is described as the formula (2):

$$\Pr(x, \lambda) = \frac{1}{2\lambda} \cdot e^{-\frac{|x|}{\lambda}}, \quad (2)$$

where  $\lambda = \frac{\Delta Q}{\epsilon}$ . The added noises are independent from the data set and are only related to the function sensitivity and the privacy parameter. The main idea of their method adds the noises subjecting to the La-



place's distribution into the output result so as to distort the sensitive data. For example, let  $Q(D)$  be the querying function of top- $k$  accessing count, then the output of the algorithm  $M$  can be represented by the following formula (3):

$$M(Q, D) = Q(D) + \left( Lap_1\left(\frac{\Delta Q}{\varepsilon}\right), Lap_2\left(\frac{\Delta Q}{\varepsilon}\right), \dots, Lap_k\left(\frac{\Delta Q}{\varepsilon}\right) \right), \quad (3)$$

where  $Lap_i\left(\frac{\Delta Q}{\varepsilon}\right)$  ( $1 \leq i \leq k$ ) is each round of the independent noise subjecting to the Laplace's distribution, and the noise is proportional to  $\Delta Q$  and inversely proportional to  $\varepsilon$ .

*Exponential mechanism:* Mcsherry and Talwar [31] described another privacy protection method, which is based on the exponential mechanism. In the method, the input of the algorithm  $M$  is the data set  $D$ , the output of the algorithm  $M$  is the result  $r$  which is selected from the output set  $Range(M)$  with the probability  $Pr$ , the function  $g: (D \times Range(M)) \rightarrow R$  is the utility measure function of the result  $r$  and  $\Delta g$  is the sensitivity of  $g$ . If the algorithm  $M$  selects the result  $r$  from  $Range(M)$  whose value is proportional to the probability  $Pr = \frac{e^{g(D,r)}}{2 \cdot \Delta g}$ , then the algorithm  $M$  satisfies  $\varepsilon$ -differential privacy.

### 3.2. Related Definitions and Theorems

**Definition 3.1.  $\varepsilon$ -Differential Privacy:** Given two adjacent data sets  $D$  and  $D'$  where at most a data record is different between  $D$  and  $D'$  ( $|D \neq D'| = 1$ ), for any algorithm  $M$  whose output range is  $Range(M)$ , if the result  $S$  outputted by the algorithm  $M$  satisfies the following formula (4) on the two adjacent data sets  $D$  and  $D'$  ( $S \in Range(M)$ ), then the algorithm  $M$  satisfies  $\varepsilon$ -differential privacy:

$$Pr[M(D) \in S] \leq e^\varepsilon \cdot Pr[M(D') \in S], \quad (4)$$

where  $Pr$  represents the randomness of the algorithm  $M$  on  $D$  and  $D'$ , namely  $Pr$  denotes the risk probability of privacy disclosure;  $\varepsilon$  represents the privacy protection level, where if  $\varepsilon$  is bigger, then privacy protection degree is lower, otherwise privacy protection degree is higher.

**Definition 3.2. Data Sensitivity<sup>2</sup>:** Data sensitivity is divided to global sensitivity and local sensitivity. Given a query function  $Q$ , the global sensitivity of the function  $Q$  is defined as follows:

$$\Delta Q = \max_{D, D'} \{|Q(D) - Q(D')|_1\}, \quad (5)$$

where  $D$  and  $D'$  represent the adjacent data sets,  $Q(D)$  represents the output of the function  $Q$  on the data set  $D$ ,  $\Delta Q$  is the sensitivity which represents the maximum of the outputs' difference.

Additionally, because the  $\varepsilon$ -differential privacy protection scheme may be used many times in the different stages of processing data, the  $\varepsilon$ -differential privacy protection scheme also needs to satisfy the following theorems:

**Theorem 3.1.** For the same data set, if the whole privacy protection process is divided to the different privacy protection algorithms ( $M_1, M_2, \dots, M_n$ ) whose privacy protection levels are  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ , then the privacy protection level  $\sum_{i=1}^n \varepsilon_i$  of the whole process needs to satisfy differential privacy protection.

**Theorem 3.2.** For the disjoint data set, if the whole privacy protection process is divided to the different privacy protection algorithms ( $M_1, M_2, \dots, M_n$ ) whose privacy protection levels are  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ , then the privacy protection level  $\max\{\varepsilon_i\}$  of the whole process needs to satisfy differential privacy protection.

## 4. Location Data Privacy Protection Scheme

In this section, we propose a location data record privacy protection scheme, which employs the structure of multi-level query tree to query and publish the data result. In the proposed scheme, we first construct the structure of multi-level query tree from data-base, and then we make double processes of selecting data

<sup>2</sup> Differential privacy protection is to add noises to protect data. If data sensitivity is small, then it can effectively protect data while a small quantity of noises are added into original data. On the contrary, if data sensitivity is big, then a lot of noises need to be added into original data.

according to access frequencies by the exponential mechanism and one process of adding noises to access frequencies by the Laplace's mechanism on the multi-level query tree. In the double processes of selecting data, the first selection is based on accessing frequency (or support count), where  $n$  location data records whose support counts are greater than a specified value are selected from the multi-level query tree; the second selection is based on the exponential mechanism where  $k$  location data records are selected from the  $n$  location data records. In the process of adding noises, noises are added into the access frequencies of the  $k$  location data records by the Laplace's mechanism. Because the multi-level query tree can satisfy the relationship of transaction data, it can meet data privacy protection requirement and data availability requirement. The procedure of the proposed scheme is described as follows (and as Table 1):

- 1 Input the data set  $D$  and the differential privacy protection parameters  $\varepsilon_1$ ,  $\varepsilon_2$ ,  $k$ ,  $min\_count$ , where  $\varepsilon = \varepsilon_1 + \varepsilon_2^3$ ;
- 2 Based on the data set  $D$  and the item set  $I$ , construct the multi-level query tree  $F_D^I$  (see Section 4.1 for more details);
- 3 Get the accessing frequency item set  $A$  from  $F_D^I$ , which satisfies that the accessing frequency of every data record is not less than  $min\_count$  in  $A^4$ ;
- 4 Use the exponential mechanism to select the accessing frequency item set  $B$  from the set  $A$ , where every selected data record satisfies the following condition:

$$\Pr(a_i) \propto \exp\left(\frac{\varepsilon_i \cdot Rank(A, a_i)}{2 \cdot \Delta Rank}\right),$$

where the size of  $B$  is  $k$ ,  $a_i \in A$  is the accessing frequency item record,  $\varepsilon_i$  is the corresponding privacy protection level,  $Rank(A, a_i)$  is the scoring value for  $a_i$  and  $\Delta Rank$  is the scoring function sensitivity (see Section 4.2 for more details);

3  $\varepsilon$  represents the privacy protection level of the whole scheme.  $\varepsilon_1$  and  $\varepsilon_2$  are independent, they are respectively used in the proposed scheme.

4 Our proposed scheme focuses on protecting highly frequent accessing location data by distorting accessing frequencies. Thus, the setting of  $min\_count$  is to improve the efficiency of the proposed scheme.

- 5 Use the Laplace's mechanism to add the noises  $Lap\left(\frac{k \cdot \Delta Q}{\varepsilon_2}\right)$  into the set  $B$ , generate the set  $C$ , and then construct and publish the new multi-level query tree according to  $C$  and  $F_D^I$  (see Section 4.3 for more details).

**Table 1**

Location data record privacy protection algorithm

---

**Input:** data set  $D$ , differential privacy protection parameters  $\varepsilon_1, \varepsilon_2, k, min\_count$   
**Output:** multi-level query tree containing noises  
**Begin**  
 Compute  $\varepsilon = \varepsilon_1 + \varepsilon_2$ ;  
 $F_D^I = \text{Construct\_Query\_Tree}(root, D, I)$ ;  
 /\*root is the root of the query tree and I is the item set\*/  
 $A = \text{Select\_Item\_Set}(F_D^I, min\_count)$ ;  
 /\*accessing frequency  $\geq min\_count$ \*/  
 $B = \text{Select\_top-}k \text{ Item\_Set}(A, \varepsilon_1)$ ;  
 /\*selection according to  $Rank()$ \*/  
 $C = \text{Add\_Laplace\_Noise}(B, \varepsilon_2)$ ;  
 $\text{Publish\_Query\_Tree}(C, F_D^I)$ ;  
**End**

---

#### 4.1. Multi-level Query Tree

This section describes how to construct a complete multi-level query tree from transaction database. As the multi-level query tree can optimize data representation, it is used to represent location data items and their accessing frequencies (counts) in this paper. Figure 1 shows that a part of the New York city is selected as the location data source.

**Figure 1**

Location data source



According to Figure 1, we set  $I = \{1, 2, 3, 4, \dots, m\}$ , where  $I$  is a set of location data items (the representation of

location data items is shown in Table 2), and set that the transaction database  $D$  is a set of location data records which include an identifier called  $T\_ID$ , the item content index  $T$  and the corresponding support (accessing) count<sup>5</sup>, where the item index  $T$  of every record is a subset of  $I$  ( $T \subseteq I$ ). The transaction database  $D$  is described in Table 3<sup>6</sup>.

**Table 2**

Representation of location data item

Item Index	Item Content
1	New York Sports Clubs
2	United States Postal Service
3	Club Quarters Hotels
4	Grand Central Market
5	Simpson Thacher and Bartlett LLP
...	...

**Table 3**

Transaction database

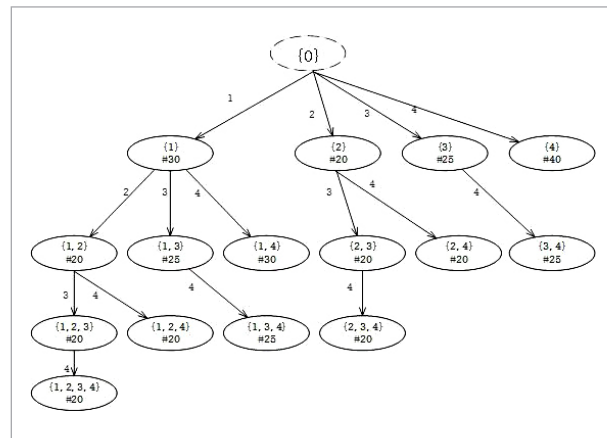
T_ID	Item Content Index	Supprt Count
T_1-T_30	1	30
T_31-T_50	2	20
T_51-T_75	3	25
T_76-T_115	4	40
T_116-T_135	1,2	20
T_136-T_160	1,3	25
T_161-T_190	1,4	30
T_191-T_120	2,3	20
...	...	...

Based on the transaction database  $D$ , we construct the multi-level query tree  $F_D^I$  as shown in Figure 2. Figure 2 shows the structure of the multi-level query tree  $F_D^I$  covering all records of the transaction database

$D$ , where  $I = \{1, 2, 3, 4\}$  is a set of location data items. In  $F_D^I$ , the root node only contains the set  $\{0\}$ , and other every node contains a label (a subset of  $I$ ) and a support (accessing) count<sup>7</sup>. For example, a label is  $\{1, 2\} \in I$ , whose accessing count is 20. The example denotes that a person accessed the zones of  $\{1\}$  and  $\{2\}$  (New York Sports Clubs and United States Postal Service) in a certain period, where the accessing count of the person is 20. From Figure 2, we may also know the relationship of the node number of  $F_D^I$  and the size of  $I$ . We can compute that the total number of the nodes is  $2^I - 1$  (does not include the root node). For example, the total number of the nodes is  $2^4 - 1 = 15$ , when  $I = \{1, 2, 3, 4\}$ .

**Figure 2**

Multi-level query Tree



**Construct\_Query\_Tree**( $root, D, I$ ) denotes the algorithm of constructing multi-level query tree. The main steps of the algorithm are described as follows<sup>8</sup>:

**Step1:** Input the current node  $cur\_node$ , the location data item set  $I$  and the transaction database  $D$ , then finish the following steps:

- the label of  $cur\_node$  becomes the union of the label of  $cur\_nod$ 's father node<sup>9</sup> and the corresponding item index from  $I$ .
- according to the label of  $cur\_node$ , the count of  $cur\_node$  equals to the corresponding support count from  $D$ .

5 The support count denotes the count that a person accessed one or several positions in a certain period.

6 Table 3 shows the related data records that a person accessed one or several positions in a certain period. For example, the data record T\_1-T\_30 denotes a person accessed the "1" position in a certain period, whose accessing count is 30. The certain period may be seen as a fixedly long duration.

7 The symbol # represents "support (accessing) count".

8 The procedure of the algorithm is similar to breadth-first traversal.

9 For the first input,  $root$  is the  $cur\_node$ 's father node and the label of  $root$  is  $\emptyset$ .

- if the label of  $cur\_node$  is equal to  $I$ , then abort all the steps; else if the max value in the label of  $cur\_node$  is less than the max value in  $I$ , then a new node as the brother node of  $cur\_node$  is created and used as the new current node  $cur\_node$ , and repeat Step 1.
- otherwise, proceed to Step 2.

**Step2:** From the first brother node of  $cur\_node$  to the last brother node of  $cur\_node$  (or  $cur\_node$ ), if the brother node of  $cur\_node$  does not have any sons, then a new node as the son node of the brother node of  $cur\_node$  is created and used as the new current node  $cur\_node$ ; repeat Step 1.

## 4.2. Selecting Accessing Frequency Item Record Set

Firstly, we get the accessing frequency item record set  $A$  from  $F_D^I$  which satisfies that the accessing frequency of every data record is not less than  $min\_count$ . Secondly, we use the exponential mechanism to select the accessing frequency item record set  $B$  from the set  $A$ , where every selected data record satisfies the following condition:

$$\Pr(a_i) \propto \exp\left(\frac{\varepsilon_i \cdot Rank(A, a_i)}{2 \cdot \Delta Rank}\right),$$

where the size of  $B$  is  $k$ ,  $a_i \in A$  is the accessing frequency item record,  $\varepsilon_i$  is the corresponding privacy protection level,  $Rank(A, a_i)$  is the scoring value for  $a_i$  and  $\Delta Rank$  is the scoring function sensitivity. The exponential mechanism can optimize the output results [31].

**Select\_top-k\_Item\_Set**( $A, \varepsilon_i$ ) denotes the algorithm of selecting accessing frequency item record set based on the exponential mechanism. The main steps of the algorithm are described as follows:

**Step1:** Input the accessing frequency item record set  $A$  where the size of  $A$  is  $N$ . Then use the scoring function  $Rank(A, a_i)$  to mark every accessing frequency item record  $a_i$  where  $a_i \in A$ ;

**Step2:** Compute the weight of every accessing frequency item record  $a_i$  according the following formula:

$$a_i.weight = \exp\left(\frac{\varepsilon_i \cdot Rank(A, a_i)}{2 \cdot \Delta Rank}\right).$$

Then rank all accessing frequency item records in descending order according to the corresponding weights, where we set  $\varepsilon_i = \frac{\varepsilon_1}{k}$ .

**Step3:** From large probability to small probability, select the accessing frequency item record set  $B$  from  $A$ , where the probability may be computed as follows:

$$\Pr(a_i) = \frac{a_i.weight}{\sum_{j=1}^N (a_j.weight)},$$

and the size of  $B$  is  $k$  ( $k \leq N$ ).

The key of selecting accessing frequency item record set is how to set the scoring function  $Rank(A, a_i)$ . In this paper, we set that  $Rank(A, a_i)$  is the support count of the corresponding node of  $a_i$  in  $F_D^I$ .  $Rank(A, a_i)$  is described as follows:

$$Rank(A, a_i) = a_i.Node.SC,$$

where  $a_i.Node.SC$  is the support count of the corresponding node of  $a_i$  in  $F_D^I$ . Then we may set the scoring function sensitivity  $\Delta Rank$  by the following method:

$$\Delta Rank = \begin{cases} k, & \text{if } k \geq \max\{Rank(A, a_i)\} - min\_count; \\ \max\{Rank(A, a_i)\} - min\_count, & \text{if } k < \max\{Rank(A, a_i)\} - min\_count, \end{cases}$$

According to the exponential mechanism, the selected probability of  $a_i$  can be calculated as follows:

$$\Pr(a_i) = \frac{\exp\left(\frac{\varepsilon_i \cdot Rank(A, a_i)}{2 \cdot \Delta Rank}\right)}{\sum_{a'_i \in A} \exp\left(\frac{\varepsilon_i \cdot Rank(A, a'_i)}{2 \cdot \Delta Rank}\right)}, \quad (6)$$

where  $\varepsilon_i$  is the corresponding privacy protection level. Because the size of  $B$  is  $k$ , we need to finish the selection of  $B$  at  $k$  times and set the value of  $\varepsilon_i$  in every round of selection. According to Theorem 3.1, we may know the privacy protection level  $\sum_{i=1}^k \varepsilon_i$  of the whole process can satisfy differential privacy protection (the detailed proof is given in Section 5).

## 4.3. Location Data Privacy Protection Algorithm Based on the Laplace's Mechanism

In this section, we show how to use the Laplace's mechanism to add the noise  $Lap\left(\frac{k \cdot \Delta Q}{\varepsilon_2}\right)$  into the set  $B^{10}$ .

<sup>10</sup>  $\Delta Q$  is the sensitivity of the query function  $Q$ , where we set  $\Delta Q = \max\{Rank(A, a_i)\} - min\_count$ .



The main steps of the algorithm are described as follows:

**Step1:** Input the privacy protection level  $\varepsilon_2$  and the accessing frequency item record set  $B$ . Then generate the noise  $Lap\left(\frac{k \cdot \Delta Q}{\varepsilon_2}\right)$  satisfying the probability  $\Pr(x, \lambda)$ , where

$$\Pr(x, \lambda) = \frac{1}{2 \cdot \lambda} e^{-\frac{|x|}{\lambda}}.$$

In the above formula, the variant  $x$  denotes the corresponding accessing frequency and  $\lambda = \frac{k \cdot \Delta Q}{\varepsilon_2}$ .

**Step2:** Add the noise  $Lap\left(\frac{k \cdot \Delta Q}{\varepsilon_2}\right)$  into the set  $B$  so as to distort the true accessing frequency of every data record by the following formula:

$$b_i \cdot SC = b_i \cdot SC + Lap\left(\frac{k \cdot \Delta Q}{\lambda}\right),$$

where  $b_i \in B$ ,  $b_i \cdot SC$  denotes the accessing frequency of  $b_i$ , and  $Lap\left(\frac{k \cdot \Delta Q}{\varepsilon_2}\right)$  is the independent noise subjecting to the probability  $\Pr(x, \lambda)$  in every round.

**Step3:** Generate the set  $C$ , then construct and publish the new multi-level query tree according to  $C$  and  $F_D^I$ .

## 5. Security Analysis of the Proposed Scheme

**Theorem 5.1.** The algorithm **Select\_top-k\_Item Set**( $A, \varepsilon_1$ ) can satisfy  $\varepsilon_1$ -differential privacy protection.

**Proof:**  $\varepsilon_1$  is the corresponding privacy protection level for the algorithm **Select\_top-k\_Item Set**( $A, \varepsilon_1$ ). We take average for  $k$  rounds of selection, namely  $\varepsilon_i = \frac{\varepsilon_1}{k}$  is the corresponding privacy protection level for every round of selection. We also set that  $S(A, j)$  is the  $j$ -th operation of selecting  $a_j$  from  $A$  with  $1 \leq j \leq k$ ,  $A'$  is the adjacent and next operated data set of  $A$  ( $|A \neq A'|=1$ ). Then, according to Definition 3.1 and Formulas (5) and (6), we may get that<sup>11</sup>

$$11 \text{ Because } \exp\left(\frac{\varepsilon_1}{2k}\right) \geq 1 \text{ and } \frac{\sum_{a_x \in A'} \exp\left(\frac{\varepsilon_1 \cdot Rank(A', a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)} \leq 1 \text{ (} A' \text{ is the}$$

adjacent and next operated data set of  $A$  with  $|A \neq A'|=1$ ), we can make the following change to the proof.

$$\begin{aligned} \frac{\Pr(S(A, j)=a_j)}{\Pr(S(A', i)=a_i)} &= \frac{\frac{\exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_j)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}}{\frac{\exp\left(\frac{\varepsilon_1 \cdot Rank(A', a_i)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A'} \exp\left(\frac{\varepsilon_1 \cdot Rank(A', a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}} \\ &= \exp\left(\frac{\varepsilon_1 \cdot (Rank(A, a_j) - Rank(A', a_i))}{k \cdot 2 \cdot \Delta Rank}\right) \cdot \left(\frac{\sum_{a_x \in A'} \exp\left(\frac{\varepsilon_1 \cdot Rank(A', a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}\right) \\ &\leq \exp\left(\frac{\varepsilon_1 \cdot \Delta Rank}{k \cdot 2 \cdot \Delta Rank}\right) \cdot \left(\frac{\sum_{a_x \in A'} \exp\left(\frac{\varepsilon_1 \cdot Rank(A', a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}\right) \\ &\leq \exp\left(\frac{\varepsilon_1}{2k}\right) \cdot \exp\left(\frac{\varepsilon_1}{2k}\right) \cdot \left(\frac{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}{\sum_{a_x \in A} \exp\left(\frac{\varepsilon_1 \cdot Rank(A, a_x)}{k \cdot 2 \cdot \Delta Rank}\right)}\right) = \exp\left(\frac{\varepsilon_1}{k}\right) \end{aligned}$$

namely,  $\frac{\Pr(S(A, j)=a_j)}{\Pr(S(A', i)=a_i)} \leq \exp\left(\frac{\varepsilon_1}{k}\right)$ , so

$$\Pr(S(A, j) = a_j) \leq \exp\left(\frac{\varepsilon_1}{k}\right) \cdot \Pr(S(A', i) = a_i).$$

Therefore, every round of selection can satisfy  $\frac{\varepsilon_1}{k}$ -differential privacy protection. According to Theorem 3.1, we may know the privacy protection level  $\varepsilon_1 = \sum_{i=1}^k \frac{\varepsilon_1}{k}$  of the whole algorithm **Select\_top-k\_Item Set**( $A, \varepsilon_1$ ) can satisfy differential privacy protection.

**Theorem 5.2.** The location data privacy protection algorithm based on the Laplace's mechanism can satisfy  $\varepsilon_2$ -differential privacy protection.

**Proof:**  $\varepsilon_2$  is the corresponding privacy protection level for the location data privacy protection algorithm. We also take average for  $k$  rounds of operation, namely  $\frac{\varepsilon_2}{k}$  is the corresponding privacy protection level for every round of operation. We set that  $spt(B, b_i)$  is the support count (accessing frequency) of  $b_i$  and  $spt(B, b_j)$  is the support count of  $b_j$  with  $b_i, b_j \in B$  and  $1 \leq i, j \leq k$ ,  $SPT$  is the support count set and  $B'$  is the adjacent data set of  $B$  ( $|B \neq B'|=1$ ). Then, according to Definition 3.1 and Formula (1), we can get that<sup>12</sup>

<sup>12</sup> In the proof, because  $|B \neq B'|=1$ , we set  $\Delta Q = \max\{Rank(A, a_i)\} - \min\_count$  in the Formula (1).

$$\begin{aligned} \frac{\Pr(spt(B,b) \in SPT)}{\Pr(spt(B',b) \in SPT)} &= \frac{\exp\left(\frac{\varepsilon_2 \cdot |spt(B,b) - spt(B,b)|}{k \cdot \Delta Q}\right)}{\exp\left(\frac{\varepsilon_2 \cdot |spt(B',b) - spt(B,b)|}{k \cdot \Delta Q}\right)} \\ &= \exp\left(\frac{\varepsilon_2 \cdot |spt(B,b) - spt(B,b)| - \varepsilon_2 \cdot |spt(B',b) - spt(B,b)|}{k \cdot \Delta Q}\right) \\ &\leq \exp\left(\frac{\varepsilon_2 \cdot |spt(B,b) - spt(B',b)|}{k \cdot \Delta Q}\right) \leq \exp\left(\frac{\varepsilon_2}{k}\right). \end{aligned}$$

Therefore, every round of operation can satisfy  $\frac{\varepsilon_2}{k}$ -differential privacy protection. According to Theorem 3.1, we may know the privacy protection level  $\varepsilon_2 = \sum_{i=1}^k \frac{\varepsilon_2}{k}$  of the whole algorithm can satisfy differential privacy protection.

## 6. Experiment Analysis of the Proposed Scheme

In this section, our experiments are mainly from two aspects to evaluate the efficiency of the proposed scheme. The first one is the running time of the proposed algorithms, which includes the time of constructing and updating the multi-level query tree and the time of extracting and protecting the available data from the multi-level query tree. The second one is protection effectiveness, which includes data utility and data privacy protection degree, where we evaluate the availability of the extracted and protected data mainly through the comparisons of true positive rate, false positive rate, accurate rate and false reject rate before and after extracting and protection.

### 6.1. Running Time Analysis

In this section, we test the running time of the proposed algorithms mainly through the time of constructing and updating the multi-level query tree and the time of extracting and protecting the available data from the multi-level query tree. All the proposed algorithms are coded by the C++<sup>13</sup> programming language. The test original data set comes from the simulation on the Baidu map<sup>14</sup>, which is similar to the

<sup>13</sup> The test environment is under Win10 OS, Intel i5 CPU 2.3Ghz and 8G RAM.

<sup>14</sup> Baidu is a network company in China. The baidu map is one of the network services provided by the company, which provides a lot of APIs for programmers to develop their applications on map.

Gowalla data set<sup>15</sup>. The test original data set contains user id, accessing time, longitude and latitude and so on. The period of the test original data set is about one month. Then we make some processes to the test original data set: 1) in the whole selected map, we select 9 zones as our tested accessing item contents (as the example of Table 2); 2) we build a table in database, whose attributes contain id, user id, accessing time and accessing item content; 3) if the longitude and latitude of the record from the test original data set is within the scope of one of the 9 zones, then the database builds one new record according to the corresponding user id in the table. Based on our experiments, the time of constructing the multi-level query tree is about 0.0014 second per 500 records, the efficiency of updating the multi-level query tree is about 10219 records per second. Table 4 shows the efficiency of extracting and protecting the available data. Table 4 shows the different numbers of extracting and protecting the available data from the multi-level query tree in a second by setting the privacy parameter  $\varepsilon = 0.01, 0.05, 0.1, 0.5, 1.1, 1.5$ , respectively.

**Table 4**

The efficiency of extracting and protecting available data

$\varepsilon$	0.01	0.05	0.1	0.5	1.1	1.5
$k$	3012	4182	40124	9131	51320	81301

From the experiments, we find that the time of constructing and updating the multi-level query tree is very fast, and the efficiency of extracting and protecting the available data from the multi-level query tree is always increasing with the increasing of  $\varepsilon$  in a certain range.

### 6.2. Protection Effectiveness Analysis

In this section, we firstly show data utility and privacy protection degree by Figures 3-5. The figures show the number change of the extracted and protected data points before and after extracting and protection. From the figures, we see that more selected data points are added around the sensitive data points after employing our proposed scheme to protect the sensitive data points, where the sensitive data can be effectively hidden and protected. Secondly, we also evaluate the availability of the extracted and protect-

<sup>15</sup> Gowalla is a location-based social networking website where users share their locations by checking-in.

ed data mainly through the comparisons of true positive rate, false positive rate, accurate rate and false reject rate before and after extracting and protection. Tables 5 and 6 show the comparisons.

Firstly, in our experiments, we set the selected record number  $k=50$ , 100 and 200 for the extracted and protected data in Figures 3-5, respectively. In the figures, the point data are the combinations of location data and accessing frequency (the point data are mapped into the figures), where the blue points represent the non-sensitive data points and the red points represent the sensitive data points. Additionally, we set the access frequency  $h$  of the extracted and protected data as the boundary of non-sensitive data and sensitive data<sup>16</sup>. The following figures show the number change of the sensitive data before and after extracting and protection.

Figure 3 shows the number change of the sensitive data before and after extracting and protection when  $k=50$ . Figure 3(a) shows that there are 8 red points representing the sensitive data before protection. Figure 3(b) shows that there are 15 red points representing the sensitive data after employing our proposed scheme, which can be seen to make the number of sensitive data points increase. Hence, the true sensitive data can be effectively hidden and protected so that the attackers cannot easily find the true sensitive data. Additionally, such a number change can also guarantee data utility, otherwise the large number change of the sensitive data points may result in data mining (or data analysis) becomes insignificant.

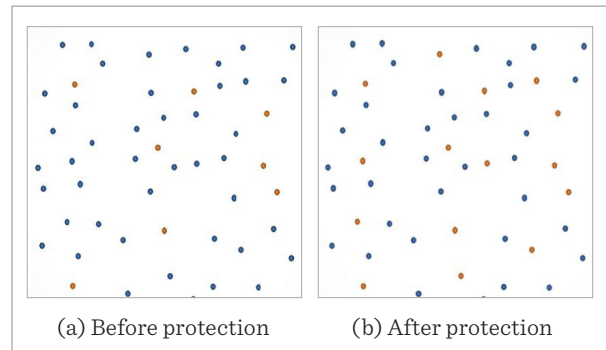
Figure 4 shows the number change of the sensitive data before and after extracting and protection when  $k=100$ . Figure 4(a) shows that there are 12 red points representing the sensitive data before protection. Figure 4(b) shows that there are 26 red points representing the sensitive data after employing our proposed scheme, which also makes the number of the sensitive data points increase.

Figure 5 shows the number change of the sensitive data before and after extracting and protection when  $k=200$ . Figure 5(a) shows that there are 15 red points representing the sensitive data before protection. Figure 5(b) shows that there are 39 red points representing the sensitive data after employing our pro-

<sup>16</sup> If the access frequency of data is more than  $h$ , then data are sensitive, otherwise they are non-sensitive.

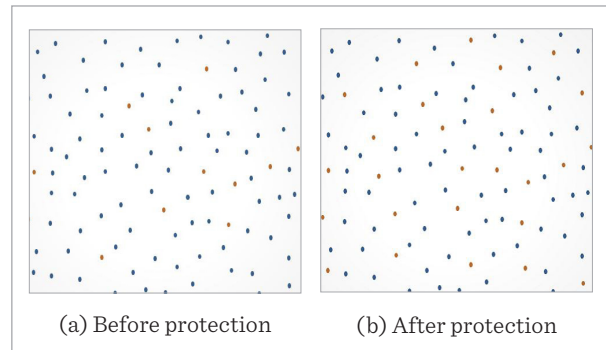
**Figure 3**

The number change of the sensitive data when  $k=50$



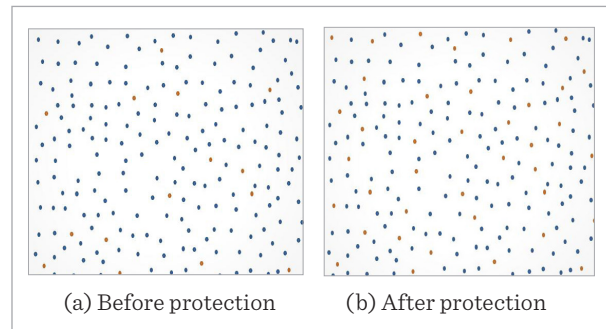
**Figure 4**

The number change of the sensitive data when  $k=100$



**Figure 5**

The number change of the sensitive data when  $k=200$



posed scheme. It denotes that the number of the red points is increased.

Secondly, we define true positive rate, false positive rate, accurate rate and false reject rate for the availability of the extracted and protected data before and after extracting and protection. Given a positive inte-

ger  $k$ , we define  $QS_k(D)$  to be the top- $k$  data set from the original multi-level query tree  $F_D^I$ , which satisfies that the accessing frequency of every data record is not less than  $min\_count$ . We also define  $QS_k'(D)$  to be the top- $k$  data set from the new multi-level query tree added by the Laplace's noises. Then we may analyze the availability of our proposed algorithms by false reject rate, where false reject rate stands for the rate of  $k$  data records being in  $QS_k(D)$  and not in  $QS_k'(D)$ . The definitions of true positive rate, false positive rate, accurate rate and false reject rate are as follows:

- 1 True positive rate: the rate of the records being in  $QS_k(D)$  and  $QS_k'(D)$ ,

$$TPR = |QS_k(D) \cap QS_k'(D)|.$$

- 2 False positive rate: the rate of the records being in  $QS_k'(D)$  and not in  $QS_k(D)$ ,

$$FPR = |QS_k'(D) - QS_k(D) \cap QS_k'(D)|.$$

- 3 Accurate rate:

$$ACY = \frac{|QS_k(D) \cap QS_k'(D)|}{|QS_k'(D)|}.$$

- 4 False reject rate:

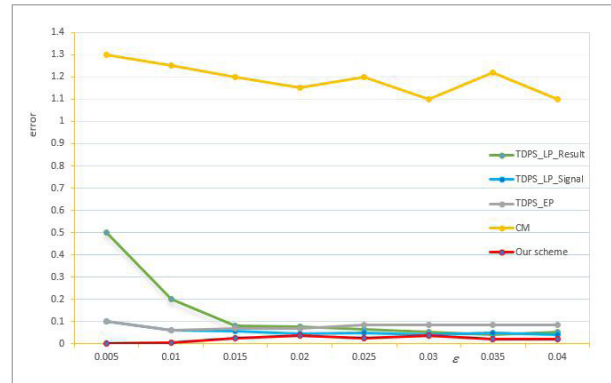
$$FRR = \frac{|QS_k(D) \cup QS_k'(D) - QS_k(D)|}{k}.$$

Therefore, we may compute error value from true positive rate, false positive rate, accurate rate and false reject rate so as to evaluate the difference of the protected data and the original data. To show the efficiency of our proposed method, we compare the **TDPS\_LP\_Result** method [32], the **TDPS\_LP\_Signal** method [32], the **TDPS\_EP** method [32] and the **CM** method [44] with our proposed method.

Figure 6 shows the error value comparison of the five methods on the noise-added data and the original data. In Figure 6, with the increase of the privacy parameter  $\varepsilon$ , the error values of the methods are becoming smaller, where the error value of our proposed method tends to be stable and very small<sup>17</sup> when  $\varepsilon > 0.015$ . Figure 6 shows that our proposed method is superior

**Figure 6**

The error comparison of the five schemes



to the other four methods. Although the error values of the **TDPS\_LP\_Result** method, the **TDPS\_LP\_Signal** method and the **TDPS\_EP** method also tend to be stable, the error values of these methods are larger than that of our proposed method. Furthermore, the error value of the **CM** method is unstable and obviously more than those of the other four methods.

In the following experiments, we compare our proposed method with the **TDPS\_LP\_Result** method, the **TDPS\_LP\_Signal** method and the **TDPS\_EP** method<sup>18</sup> on true positive rate, false positive rate and accurate rate; and we compare our proposed method with the **DP-top- $k$**  method [15] on false reject rate.

When we set the privacy parameter to  $\varepsilon=1.1$ , Table 5 shows the comparisons of our proposed method, the **TDPS\_LP\_Result** method, the **TDPS\_LP\_Signal** method and the **TDPS\_EP** method on true positive rate, false positive rate and accurate rate, where  $k=20, 40, 60, 80, 100$  and  $200$ . In Table 5, with the increase of  $k$ , the accurate rates of the four methods decrease. When  $k=200$ , the accurate rates tend to be stable, where the accurate rates of the **TDPS\_LP\_Result** method and the **TDPS\_LP\_Signal** method are about 63%, the accurate rate of the **TDPS\_EP** method is below 35% and the accurate rate of our proposed method is stably about 80%. Therefore, even if the value of  $k$  becomes big, our proposed method can also maintain high accurate rate.

<sup>18</sup> In Tables 5 and 6, the symbol R denotes the **TDPS\_LP\_Result** method, the symbol S denotes the **TDPS\_LP\_Signal** method, the symbol E denotes the **TDPS\_EP** method, the symbol O denotes our proposed method.

<sup>17</sup>  $\varepsilon$  is smaller, privacy protection level is higher.



**Table 5**

The data efficiency comparison of the four schemes

<i>k</i>	TPR				FPR				ACY			
	R	S	E	O	R	S	E	O	R	S	E	O
20	19	20	17	20	1	0	3	0	0.95	1	0.85	1
40	38	38	22	39	2	2	18	1	0.95	0.95	0.55	0.98
60	54	57	30	58	6	3	30	2	0.9	0.95	0.5	0.97
80	67	66	34	73	13	14	46	7	0.84	0.83	0.425	0.91
100	80	78	39	90	20	22	61	10	0.8	0.78	0.39	0.9
120	90	92	46	105	30	38	74	15	0.75	0.77	0.38	0.88
140	94	101	52	119	46	39	88	21	0.67	0.72	0.371	0.83
160	108	109	60	128	52	51	100	32	0.68	0.68	0.375	0.8
180	117	117	65	144	63	63	115	32	0.65	0.65	0.36	0.8
200	126	125	70	160	74	75	130	40	0.63	0.63	0.35	0.8

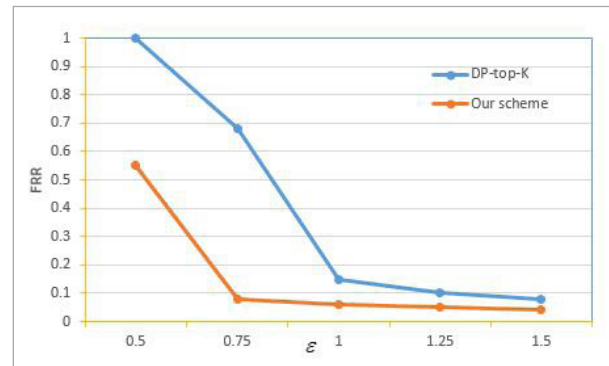
Additionally, when we fixedly set  $k=60$ , we analyzed the comparisons of our proposed method, the **TDPS\_LP\_Result** method and the **TDPS\_LP\_Signal** method by changing the value of  $\epsilon$  where  $\epsilon=0.01, 0.05, 0.1, 0.5, 1.1$  and  $1.5$ . Table 6 shows the comparisons. In Table 6, with the increase of  $\epsilon$ , the accurate rates of the three methods increase. When  $\epsilon>1.1$ , the accurate rates of the **TDPS\_LP\_Result** method and the **TDPS\_LP\_Signal** method tend to be stable above 92%, the accurate rate of our proposed method is about 97%.

In Figure 7, we also compare our proposed method with the **DP-top- $k$**  method on false reject rate. When we fixedly set  $k=100$ , we analyzed the comparisons of our proposed method and the **DP-top- $k$**  method by changing the value of  $\epsilon$  from 0.5 to 1.5. In Figure 7, the false reject rates of our proposed method and the **DP-top- $k$**  method can both maintain low values with the change of  $\epsilon$ . Further, because the false reject rate of

our proposed method can maintain lower value than that of the **DP-top- $k$**  method, our proposed method is more effective.

**Figure 7**

The FRR comparison of the two schemes when  $k=100$



**Table 6**

The data efficiency comparison of the three schemes when  $k=60$

$\epsilon$	TPR			FPR			ACY		
	R	S	O	R	S	O	R	S	O
0.01	40	56	54	20	4	6	0.67	0.93	0.90
0.05	54	57	54	6	3	6	0.90	0.95	0.90
0.1	52	56	54	8	4	6	0.87	0.93	0.90
0.5	54	55	55	6	5	5	0.90	0.97	0.92
1.1	55	57	58	5	3	4	0.92	0.95	0.97
1.5	55	57	58	5	3	2	0.92	0.95	0.97

## 7. Conclusions

As the records about location data may be discrete in database, some existing privacy protection schemes are difficult to protect location data in data mining. In this paper, we propose a location data record privacy protection scheme based on differential privacy mechanism, which employs the structure of multi-level query tree to query and publish location data on database. In the proposed scheme, we first construct the structure of multi-level query tree from database, then we make double processes of selecting data according to accessing frequencies by

the exponential mechanism and one process of adding noises to accessing frequencies by the Laplace's mechanism on the multi-level query tree. The ex-

periments show that the data availability of the proposed scheme is higher and the privacy protection of the scheme is effective.

## References

1. Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J. Google Street View: Capturing the World at Street Level. *Computer*, 2010, 43(6), 32-38. <https://doi.org/10.1109/MC.2010.170>
2. Backstrom, L., Dwork, C., Kleinberg, J. Wherefore art thouR3579X?: Anonymized Social Networks, Hidden Patterns and Structural Steganography. *Proceedings of the 16th International Conference on World Wide Web, Banff, 2007*, 181-190. <https://doi.org/10.1145/1242572.1242598>
3. Bamba, B., Liu, L., Pesti, P., Wang, T. Supporting Anonymous Location Queries in Mobile Environments with Privacy Grid. *Proceedings of the 17th International Conference on World Wide Web. New York: ACM Press, 2008*, 237-246. <https://doi.org/10.1145/1367497.1367531>
4. Beresford, A., Rice, A., Skehin, N., Sohan, R. MockDroid: Trading Privacy for Application Functionality on Smartphones. *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, ACM Press, 2011*, 49-54. <https://doi.org/10.1145/2184489.2184500>
5. Bindschaedler, V., Shokri, R. Synthesizing Plausible Privacy-Preserving Location Traces. *IEEE Symposium on Security and Privacy, 2016*, 546-563. <https://doi.org/10.1109/SP.2016.39>
6. Chatzikokolakis, K., Palamidess, C., Stronati, M. A Predictive Differentially-Private Mechanism for Mobility Traces. *Proceedings of the 14th International Symposium on Privacy Enhancing Technologies, Berlin: Springer, 2014, LNCS 8555*, 21-41. [https://doi.org/10.1007/978-3-319-08506-7\\_2](https://doi.org/10.1007/978-3-319-08506-7_2)
7. Chatzikokolakis, K., Palamidessi, C., Stronati, M. Geo-Indistinguishability: A Principled Approach to Location Privacy. *ICDCIT 2015, Berlin: Springer, 2015, LNCS 8956*, 49-72. [https://doi.org/10.1007/978-3-319-14977-6\\_4](https://doi.org/10.1007/978-3-319-14977-6_4)
8. De Cristofaro, E., Soriente, C., Tsudik, G., Williams, A. Hummingbird: Privacy at the Time of Twitter. *IEEE Symposium on Security and Privacy, San Francisco, 2012*, 285-299. <https://doi.org/10.1109/SP.2012.26>
9. Dwork, C. A Firm Foundation for Private Data Analysis. *Communications of the ACM*, 2011, 54(1), 86-95. <https://doi.org/10.1145/1866739.1866758>
10. Dwork, C., McSherry, F., Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. *Proceedings of the 3th Theory of Cryptography Conference (TCC06), Berlin: Springer, 2006*, 265-284. [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14)
11. Dwork, C. The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques. *Proceedings of the Foundations of Computer Science (FOCS), Piscataway, NJ: IEEE, 2011*, 1-2. <https://doi.org/10.1109/FOCS.2011.88>
12. FireEagle. Available at: <http://info.yahoo.com/privacy/us/yahoo/fireeagle/>.
13. Frommer, D. Loopt Location to Update in the Background on iPhone, 2009. Available at: <http://www.businessinsider.com/loopt-to-run-in-the-background-on-iphone-2009-6>.
14. Fung, B., Wang, K., Yu, P. Anonymizing Classification Data for Privacy Preservation. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(5), 711-725. <https://doi.org/10.1109/TKDE.2007.1015>
15. Gedik, B., Liu, L. Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms. *IEEE Transactions on Mobile Computing*, 2008, 7(1), 1-18. <https://doi.org/10.1109/TMC.2007.1062>
16. Glass, J. Shyhood is Location, 2014. Available at: <http://www.skyhookwireless.com/>.
17. Google Latitude. Available at: <http://www.google.com/latitude/apps/badge>.
18. He, X., Cormode, G., Machanavajhala, A., Procopiuc, CM., Srivastava, D. DPT: Differentially Private Trajectory Synthesis Using Hierarchical Reference Systems. *Proceedings of the VLDB Endowment*, 2015, 8(11), 1154-1165. <https://doi.org/10.14778/2809974.2809978>
19. Huo, Z., Meng, X. A Survey of Trajectory Privacy-Preserving Techniques. *Chinese Journal of Computers*, 2011, 34(10), 1820-1830. <https://doi.org/10.3724/SP.J.1016.2011.01820>
20. Kim, M., Fielding, J., Kotz, D. Risks of Using AP Locations Discovered Through War Driving. *International Conference on Pervasive Computing, Berlin: Springer, 2006, LNCS 3968*, 67-82. [https://doi.org/10.1007/11748625\\_5](https://doi.org/10.1007/11748625_5)

21. Korolova, A., Motwani, R., Nabar, S., Xu, Y. Link Privacy in Social Networks. Proceedings of the 24th International Conference on Data Engineering, Cancun, 2008, 1355-1357. <https://doi.org/10.1109/ICDE.2008.4497554>
22. LeFevre, K., DeWitt, D., Ramakrishnan, R. Mondrian Multidimensional k-Anonymity. Proceedings of the 22nd International Conference on Data Engineering, Atlanta, 2006, 6(3), 25-35. <https://doi.org/10.1109/ICDE.2006.101>
23. Li, N., Li, T., Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. Proceedings of the 23rd IEEE International Conference on Data Engineering, Istanbul, 2007, 106-115. <https://doi.org/10.1109/ICDE.2007.367856>
24. Li, N., Qardaji, W., Su, D., Cao, J. PrivBasis: Frequent Itemset Mining with Differential Privacy. Proceedings of the VLDB Endowment, 2012, 5(11), 1340-1351. <https://doi.org/10.14778/2350229.2350251>
25. Liu, F., Hua, K., Cai, Y. Query l-Diversity in Location-Based Services. Proceedings of the 10th International Conference on Mobile Data Management, Taipei, 2009, 436-442. <https://doi.org/10.1109/MDM.2009.72>
26. Liu, J., Cai, B., Wang, J. An Analysis of BeiDou Navigation Satellite System (BDS) Based Positioning for Train Collision Early Warning. Intelligent Vehicles Symposium, 2013, 36(1), 1065-1070. <https://doi.org/10.1109/IVS.2013.6629607>
27. Liu, L. From Data Privacy to Location Privacy: Models and Algorithms. Proceedings of the 33rd International Conference on Very Large Data Bases. New York: ACM Press, 2007, 1429-1430.
28. Li, Y., Zhang, Z., Winslett, M., Yang, Y. Compressive Mechanism: Utilizing Sparse Representation in Differential Privacy. Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society, New York: ACM, 2011, 177-182. <https://doi.org/10.1145/2046556.2046581>
29. Loki. Available at: <http://loki.com/>.
30. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramanian, M. L-Diversity: Privacy Beyond k-Anonymity. Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta, 2006. <https://doi.org/10.1109/ICDE.2006.1>
31. Mcsherry, F., Talwar, K. Mechanism Design Via Differential Privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Piscataway, NJ: IEEE, 2007, 94-103. <https://doi.org/10.1109/FOCS.2007.4389483>
32. Ouyang, J., Yin, J., Liu, S., Liu, Y. An Effective Differential Privacy Transaction Data Publication Strategy. Journal of Computer Research & Development, 2014, 51(10), 2195-2205. <https://doi.org/10.7544/issn1000-1239.2014.20130824>
33. Roberts, P., Challinor, S. IP Address Management. BT Technology Journal, 2000, 18(3), 127-136. [https://doi.org/10.1049/PBBT001E\\_ch14](https://doi.org/10.1049/PBBT001E_ch14)
34. Sadeh, N. M-Commerce: Technologies, Services, and Business Model (1st edition). Wiley: New York, 2002, 241-248.
35. Samarati, P. Protecting Respondents Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6), 1010-1027. <https://doi.org/10.1109/69.971193>
36. Samarati, P., Sweeney, L. Generalizing Data to Provide Anonymity when Disclosing Information. Proceedings of the 7th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems, Seattle, 1998, 188-202. <https://doi.org/10.1145/275487.275508>
37. Sousa, M., Techmer, A., Steinhage, A., Lauterbach, C., Lukowicz, P. Human Tracking and Identification Using a Sensitive Floor and Wearable Accelerometers. Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), San Diego, 2013, 166-171. <https://doi.org/10.1109/PerCom.2013.6526728>
38. Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5), 571-588. <https://doi.org/10.1142/S021848850200165X>
39. Tao, Y., Xiao, X., Li, J., Zhang, D. On Anti-Corruption Privacy Preserving Publication. Proceedings of the 24th International Conference on Data Engineering, Cancun, 2008, 725-734. <https://doi.org/10.1109/ICDE.2008.4497481>
40. Tsai, J., Kelley, P., Cranor, L., Sadeh, N. Location-Sharing Technologies: Privacy Risks and Controls. A Journal of Law and Policy for the Information Society, 2010, 6(2), 119-152.
41. Ugolotti, R., Sassi, F., Mordonini, M., Cagnoni, S. Multi-Sensor System for Detection and Classification of Human Activities. Journal of Ambient Intelligence and Humanized Computing, 2013, 4(1), 27-41. <https://doi.org/10.1007/s12652-011-0065-z>

42. Wang, N., Xiao, X., Yang, Y., Zhang, Z., Gu, Y., Yu, G. Priv-Super: A Superset-First Approach to Frequent Itemset Mining under Differential Privacy. *IEEE International Conference on Data Engineering*, San Diego, 2017, 809-820. <https://doi.org/10.1109/ICDE.2017.131>
43. Wang, Q., Zhang, Y., Lu, X., Wang, Z., Qin, Z., Ren, K. RescuedP: Real-Time Spatio-Temporal Crowd-Sourced Data Publishing with Differential Privacy. *IEEE INFOCOM 2016*, San Francisco, 2016, 1-9. <https://doi.org/10.1109/INFOCOM.2016.7524458>
44. Wang, T., Liu, L. *From Data Privacy to Location Privacy. Machine Learning in Cyber Trust-Security, Privacy, and Reliability*, Berlin: Springer, 2009, 217-246. [https://doi.org/10.1007/978-0-387-88735-7\\_9](https://doi.org/10.1007/978-0-387-88735-7_9)
45. Wong, R., Fu, A., Wang, K., Pei, J. Minimality Attack in Privacy Preserving Data Publishing. *Proceedings of the 33rd International Conference on Very Large Data Bases*, Vienna, 2007, 543-554.
46. Wong, R., Li, J., Fu, A., Wang, K. ( $\alpha$ ,  $k$ )-Anonymity: An Enhanced  $k$ -Anonymity Model for Privacy-Preserving Data Publishing. *Proc. of the ACM 12th SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 2006, 754-759. <https://doi.org/10.1145/1150402.1150499>
47. Xiao, X., Tao, Y. Anatomy: Simple and Effective Privacy Preservation. *Proceedings of the 32nd International Conference on Very Large Data Bases*, New York: ACM Press, 2006, 139-150.
48. Xiao, X., Tao, Y. Personalized Privacy Preservation. *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, Chicago, 2006, 229-240. <https://doi.org/10.1145/1142473.1142500>
49. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A. Utility-Based Anonymization Using Local Recoding. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 2006, 785-790. <https://doi.org/10.1145/1150402.1150504>
50. Zhang, Q., Koudas, N., Srivastava, D., Yu, T. Aggregate Query Answering on Anonymized Tables. *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*, Piscataway, NJ: IEEE, 2007, 116-125. <https://doi.org/10.1109/ICDE.2007.367857>
51. Zhang, X., Wang, M., Meng, X. An Accurate Method for Mining Top- $k$  Frequent Pattern Under Differential Privacy. *Journal of Computer Research and Development*, 2014, 51(1), 104-114. <https://doi.org/10.7544/issn1000-1239.2014.20130685>
52. Zheleva, E., Getoor, L. Preserving the Privacy of Sensitive Relationships in Graph Data. *Proceedings of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD*, San Jose, 2007, 153-171. [https://doi.org/10.1007/978-3-540-78478-4\\_9](https://doi.org/10.1007/978-3-540-78478-4_9)