Privacy Protection Method on Publishing Dynamic Set-Valued Data

Jian Zhang and Yu Yang
Department of Information Engineering, Engineering University of CAPF, Xi'an, Shanxi, China

***Abstract*—After researching many protection methods for releasing sensitive information, we found that differential privacy could provide strong guarantees on it, but it will adding too much noise and spending too much times in releasing dynamic set-valued data. To solve these problems, this paper present a privacy protection method based on Diffpart. For a dataset need to be published, set it by Diffpart algorithm firstly, then using sampling method to sample some nodes and add the Laplace noise to them to protect the sensitive information when dataset needs to be updated. Then, generate a transposing number randomly to adjust the sampling nodes for subsequent update. Through the experiment, compared with the Diffpart algorithm, the method that we raised reached a ideal effect in practicability and protective for data release.**

Keywords—set-valued data; differential privacy; dynamic release; taxonomy tree; sample

I. INTRODUCTION

With the widely used of computer networks and applications, personal data has become important resources in all areas, which played an important role in the development of society. For the needs of study and share, some data will be released on the Internet, but release it directly to the dataset will cause leakage risks of privacy information such as personal name, disease, and shopping information, which could cause potential harm to individuals, or enterprise. Therefore, the protection of the privacy in data release becomes a key point in the field of information security in today's society.

In order to release the interactive data better, after researching many methods of data publication, found that taking a top-down taxonomy tree, combined with differential privacy method can effectively protect the data privacy information. In many taxonomy tree methods [1], Diffpart algorithm is one of the most representative ways [2]. this paper adopt differential privacy mechanism, combined with the feature of set-valued data, and the characteristic of the dynamic data release, DR-Diff algorithm is proposed, which improves the add method of noise, and make it feasible to release dynamic data. In release data for the first time, node would be divided based on Diffpart, and add noise to the node that sampled randomly. In the subsequent dynamic release, through random dynamic adjust the sampled point location, making the data digger cannot determine the location of the noise and its influence on data, so as to achieve the dynamic release, decrease noise adding, improve data availability.

II. BACKGROUND

A. Set-valued Data

Set-valued data [3], such as the shopping data, can be expressed as $\{id, \{item1, item2, \cdots, item_n\}\}$, each record consists of a code and a set of elements. Any one or a combination of several elements can be sensitive attributes, disclosure the privacy information.

B. Differential Privacy

Differential privacy [4-6] protection model was put forward by Dwork.

For two data sets D_1 and D_2, differs in the records of at most one individual. A is a random algorithm, range(A) refers to all output of A, S is subset of range(A).If the algorithm A satisfied:

$$\Pr[A(D_1) \in S] \le e^{\varepsilon} \times \Pr[A(D_2) \in S] \tag{1}$$

The algorithm has $\varepsilon-$differential privacy which has a protection budget ε, the smaller the of the ε, the higher the protection level it has.

C. Laplace Mechanism [7]

For any function $f: D \to R^d$, the mechanism

$$Y(D) = f(D) + \mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \tag{2}$$

gives $\varepsilon-$differential privacy. Among them, the $\mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right)$ is density function

$$\Delta f = \max_{D_1,D_2} |f(D_1) - f(D_2)| \tag{3}$$

As a query sensitivity of $f(D)$.

D. Diffpart Taxonomy Tree

First generate all distinct item sets from the item universe; then for each item set issue a counting query and add Laplace noise to the answer. Exactly distribution of privacy budget has great influence in the degree of secrecy.

E. Characteristic of Differential Privacy [8-9]

Sequence combination: Given data base D and random algorithms $A_1, \cdots, A_n$, and $A_i (1 \le i \le n)$ meet $\varepsilon-$differential privacy. In that case, $\{A_1, \cdots, A_n\}$ meet $\varepsilon-$differential privacy among the data base, $\varepsilon = \sum \varepsilon_i$.

Parallel combination: Given a data base D which divided into N disjoint subsets$D=\{D_1, \cdots, D_n\}$, and A is a random algorithm that meet $\varepsilon-$differential privacy. In that case the algorithm A meet $\varepsilon-$differential privacy in $\{D_1, \cdots, D_n\}$.

III. FUNDAMENTAL OF DR-DIFF

Under the environment of big data, data set contains a variety of data items, and need to be continually updated. This leads to the leaf node number too large in top-down iterative segmentation process in Diffpart algorithm, if add noise to each of the nonempty nodes, could cause problems such as waste and shortage of privacy budget. When publishing dynamic data over time, the updated data set must contain noises added to the previous released data set, which will cause a huge difference with the accumulation of noise with the data release for many times, reducing the authenticity and availability of data.

The study found that by using the fixed sampling technology, only adding noise to extracted node can significantly reduce the amount of noise added and the impact on the data availability. When dynamic release data, through adopts the method of random dynamic fixed sampling to determine the location of adding noise, making data digger or malicious attackers will not be able to determine which data is produced after the interference, which data is untreated, so as to achieve the protection of data availability and data privacy. When adding noise, privacy budget will be allocated to each sampling point, according to the principle of "sequence combination", ensure all the data meet the $\varepsilon-$ differential privacy.

A. Data Release Process

1) The first release process: For a data table to be released, first use Diffpart to process the data set, in order to make its reach a certain protective effect.

2) The first update release process: random choose n nodes from the prior taxonomy tree, add Laplace noise on the selected node, the data between the two nodes don't do anything, and then publish the data table.

3) The follow-up dynamic release process: After determining displacement value C, add Laplace noise to the node which shift C to the right side, data between the two nodes don't do processing, and then redistribute the data table. The algorithm process is shown in Figure I:

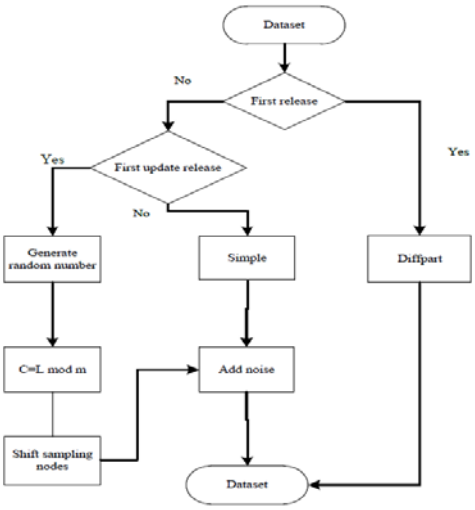

FIGURE I. FLOW CHART

B. Sampling Point Selection Method

1) Selection of Fixed Node: First of all, according to the need to determine the sampling interval t, node interval according to the actual data attribute and the degree of protection, the protection degree is higher, the smaller the node spacing. Subsequently, random select $1 \le r_1 \le \mathrm{t}$ as a starting node, the rest of the nodes are:

$$\mathrm{r}_\mathrm{i}' \in \{\mathrm{r}_\mathrm{i}' = \mathrm{r}_1 + (\mathrm{k}-1) \times \mathrm{t} | 1 \le \mathrm{i} \le \mathrm{n}, 1 < k, k \in Z\} \quad (4)$$

2) Selection of Shift Number: Set the total number of taxonomy tree's leaf nodes is m, using Matlab to produce a pseudo-random number L of length l, calculation:

$$\mathrm{c} = \mathrm{L}\ \mathrm{Mod}\ \mathrm{m} \quad (5)$$

The nodes added noise are:

$$\mathrm{r}_\mathrm{i}'' = \mathrm{r}_\mathrm{i}' + \mathrm{c} \quad (6)$$

The DR-Diff is shown in algorithm 1. It describes the basic algorithm of dynamic data release. Using sampling techniques to add noise and realize the dynamic release process, availability can be greatly improved in the process of the protection of privacy data. Experiments will prove that the use of sampling techniques can effectively protect the privacy of users in the dynamic release. Diffpart algorithm has been described in literature [2].

Algorithm 1: DR-Diff

Input: dynamic data set X, privacy budget ε, k for any point in time

Output: Data set R

a) if $k = 0$

b) Implementation Diffpart algorithm for X

c) else if $k = 1$

d) Choose sampling interval t

e) Calculate the sampling point number $n = \mathrm{INT}\left(\frac{m}{t}\right)$

f) Add noise to sampling points $r_i' = r_1 + (k-1) \times t + Lap\left(\frac{\varepsilon}{n}\right)$, $1 \le i \le n, 1 \le k, k \in Z$

g) else

h) Generate a random number $\mathrm{L} = \mathrm{INT}(1000 \times \mathrm{rand}(1,1))$

i) Calculate the displacement number $\mathrm{c} = \mathrm{Mod}(\mathrm{L}, \mathrm{m})$

j) add noise to sampling points $r_i'' = r_i' + c + Lap\left(\frac{\varepsilon}{n}\right)$

k) end

IV. EXPERIMENTAL EVALUATIONS

In the experiment, we tested the practicality of the algorithm in the different data mining tasks and usability on large data sets of set-valued data. In the experiments, we will

compare DR-Diff algorithm with Diffpart algorithm, to prove that DR-diff algorithm can reduce the amount of noise added and the influence of interference to the data, and have better practicability in dealing with a large amount of data. Meanwhile, the DR-Diff realized the application of dynamic data release.

In order to verify the algorithm's actual effect, evaluated the DR-diff from the execution time, information loss, and the average relative error of dynamic release. The implementation was done in C# ,and all experiments were conducted on an Inter Core^{TM} i5 CPU 2.67GHz ; 4G RAM; WIN7 OS. Realize the algorithm in Eclipse8.5. The experiment using MSNBC and STM two real data sets. MSNBC is a small size database which describes URL categories visited by users, we convert is to the set-valued data by ignoring its sequentiality. STM is consisted by the passengers ride records. Data statistics are shown in table I, among them, the D is the total number of records, I as the total number of element types in the records, max |t| is the maximum number of species in a record.

TABLE I. EXPERIMENTAL DATASET STATISTICS

Data Set	|D|	|I|	max|t|
MSNBC	989818	17	17
STM	1210096	1012	64

A. *Efficiency Test and Analysis*

In order to verify the DR-Diff algorithm is better than the original Laplace noise mechanism on efficiency. We randomly select 3000 experimental data from MSNBC to simplify the experiment. The efficiency of different differential privacy mechanism has nothing to do with privacy budget, Figure II shows in the comparison of different algorithms in system efficiency under the condition of $\varepsilon = 1.0$.

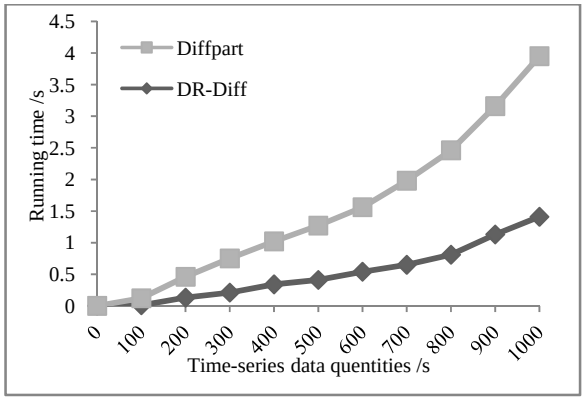

FIGURE II. EFFICIENCY OF DIFFERENT ALGORITHM

By the experimental results it can be seen that DR-Diff algorithm's running time significantly below Diffpart algorithm, this is due to the DR-Diff algorithm only to add noise to sampling point data, so has a significant advantage in running time.

B. *Degree of Privacy Protection Test and Analysis*

Experiments have carried on many times on a given budget $\varepsilon = 1.0$ and different sampling interval. Using DR-Diff algorithm to build the decision tree model, and get the release dataset and test its average relative error, the average relative error showed as negative logarithm, the smaller the value, the greater the average relative error it has. The experimental results are shown in Figure II.

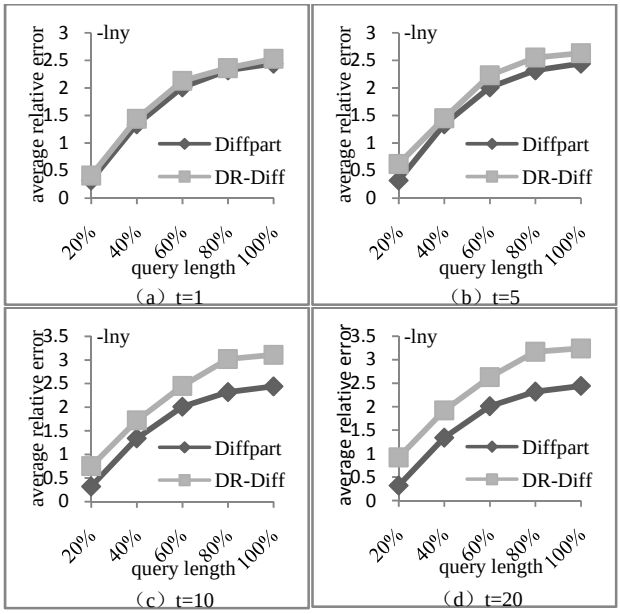

FIGURE III. RESULTS OF DIFFERENT QUERY LENGTH

From experimental results, it can be seen that DR-Diff and Diffpart algorithm has similar average relative error when the sampling interval is small, this is due to the small sampling interval basically cause the similar noise added with the original algorithm, with the increase of the sampling interval, the disturbance significantly reduced, DR-Diff has the better performance than the original algorithm in average relative error.

With the comparison with Diffpart algorithm at run time and the average relative error, it can be seen that the DR-Diff algorithm has a higher efficiency with the increment of time-series data, and has a higher availability under the condition of the sampling interval increases.

V. CONCLUSIONS

In this paper, aiming at the dynamic release of set-valued type data problem, DR-Diff algorithm is proposed. Firstly, the basic principle of differential privacy and the set-valued data are introduced. Secondly, we summarized the existing data algorithm and points out its deficiency. Finally, out forward an algorithm based on the Diffpart data publishing algorithm. And through experimental verification, the results show that the proposed algorithm has reached the ideal effect on the protective and practicability.

REFERENCES

[1] Inan A, Kantarcioglu M, Ghinita G, et al. Private record matching using differential privacy[C]//Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010: 123-134.

[2] [2] Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098.

[3] YU D, KANG H. Privacy protection method on time-series data publication[J]. Journal on Communications; 2015, 36(Z1): 243-249.

[4] Zhang X, Wang M, Meng X, An Accurate Method for Mining top-k Frequent Pattern Under Differential Privacy[J]. Journal of Computer Research and Development, 2014, 51(1): 104-114.

[5] Xiong P, Zhu T , Wang X, A Survey on Differential Privacy and Applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.

[6] Dwork C. Differential privacy: A survey of results[C]//International Conference on Theory and Applications of Models of Computation. Springer Berlin Heidelberg, 2008: 1-19.

[7] Dwork C. Differential Privacy in New Settings[C]//SODA. 2010: 174-183.

[8] Zhang X, Meng X, Chen R. Differentially private set-valued data release against incremental updates[C]//International Conference on Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2013: 392-406.

[9] McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009: 19-30.