Adaptive Differential Privacy Interactive Publishing Model Based on Dynamic Feedback

Laifeng Lu[①②] *,Yanping Li* [②(*)]
[①]*Guizhou Provincial Key Laboratory of Public Big Data，Guizhou Universtiy, Guiyang, China, 550025,*
[②]*School of Mathematics and Information Science，*
Shaanxi Normal University，Xi'an, China,710119
lulaifeng@snnu.edu.cn, lyp@snnu.edu.cn
**Corresponding author: Yanping Li*

Yihui Zhou[③]*,Feng Tian*[③]*,Hai Liu*[③]
[③]*School of Computer Science ,*
Shaanxi Normal University，Xi'an China,710119
zhouyihui@snnu.edu.cn, tianfeng@snnu.edu.cn,
liuhai @snnu.edu.cn

Abstract*-Data publishing is very meaningful and necessary. However, there are much personal (especially sometimes sensitive) information in the datasets to be published. So, privacy preserving has become a more and more important problem what we must deal with in big data era. Because of the strong mathematics foundation, provable and quantized privacy properties, DP (differential privacy) attracts the most interests and is becoming one of the most prevalent privacy models.*

This paper, based on differential privacy preserving mechanism, engages in queries restriction problem in interactive privacy data publishing framework. One adaptive differential privacy interactive publishing model based on dynamic feedback model (ADPM-DF) is proposed. Then, its technological process is presented by the flow chart in detail. And, the dynamic feedback scheme is proposed with an iteration algorithm to generate new privacy budget parameter. Finally, some qualities are discussed. Analysis shows that the new model can run well with good practical meanings and provide better user query experience.

Keywords-Differential Privacy Preserving; Interactive Publishing Framework; Laplace Mechanism; Privacy Budget; Feedback Mechanism;

I. *INTRODUCTION*

Nowadays, kinds of enterprises, companies, websites and governments keep collecting people's personal data. The datasets are from education, economy, traffic and so on. The big data era is coming. [1]. For example, the learners' study process and scores are recorded in detail by the study system online. If we go to hospital to visit a doctor, our medical information will be collected by hospitals. Once we use the search engines to search topic we focus on, or we communicate with other people on social network sites, or we buy something on the e-commerce sites, our behaviors and personal habits online are recorded. If we use GPS system, our position information and travel details are collected by the system. What's more, the speed of digital information collected keeps developing rapidly in last few years.

Data publishing (also called data release), who aims to share datasets or some query results to public users, is quite necessary [2]. Because by the means of many popular computer science technologies, such as DM(data mining), DML (deep machine learning) and etc., data publishing can bring economic and social benefits, such as offering better suggestion, recommendation and personal services, publishing much more accurate statistics information, etc.

However, most of the datasets collected by kinds of methods listed as above contain huge personal and sensitive information, such as personal identities, health information, location information and other privacies. Even *k*-anonymity[3], *l*-diversity[4], *t*-closeness[5] and some other privacy preserving techniques[6-7] are used to hide such significant privacy information, data publishing may also lead to sensitive information leakage. So, privacy preserving is becoming a more and more urgent and popular research domain in the big data era. We can adopt the "differential privacy preserving" method to deal with the privacy leakage in data publishing scenario.

Actually, DP (differential privacy)[8] is a privacy preserving framework which is new emerging and full of vitality. Because of its strong mathematics foundation, provable and quantized privacy properties, DP attracts many researchers' interests. It has been applied in wide ranges, both in theories and kinds of application scenarios. So, in this paper, we will discuss and deal with the problems during the dataset publishing based on differential privacy.

A. The Problems DPP facing

In the data publishing system based on differential privacy (DPP), the general step is as following: User sends some certain queries to data curator (datasets managers). The curator deals with the original query result with differential privacy technology (such as Laplace mechanism), that is add some random noise to some of the query results. After that, the new query result will be feedback to the query user.

However, without any restriction of the numbers of queries, the privacy leakage risk will enlarge with the enlargement of query count. How to develop privacy preserving and decrease the privacy leakage risk with the large query number? And during the process of privacy preserving, we hope it can bring much better experience to

978-1-5386-8303-3/18/$31.00 ©2018 IEEE
DOI 10.1109/NaNA2018.2018.00047

the query users. Otherwise, the simple and rough method is once the query count is larger than the limit, the user will be forbidden to query again. Some works will be expected to solve the problem above.

B. Related works

In Dwork's earlier work, Laplace mechanism was designed[8-9]. It can do well to all real-values queries. However, the total count of queries should be bounded by the size of dataset. Otherwise, with the query numbers enlargement, some records of the datasets can be deduced by the attackers.

Dinur et al. develop the research of data publishing and query numbers limitation[10]..

C. Our contributions and paper organization

In this paper, we engage in discussing privacy preserving interactive publishing query restriction problem based on differential privacy with better experience. Our contributions and the rest of this paper are developed as following:

In the second part, some preliminaries about differential privacy are introduced, such as the definition of differential privacy, privacy budget, some privacy theorems and so on. Then the third section is coming, which is the most important section of this paper. One new model *ADPM-DF*(adaptive differential privacy interactive publishing model based on dynamic feedback) is proposed. Then based on consideration of implementation, we provide the technological flow chart in detail. After that, one general feedback scheme to generate new privacy budget parameter is designed and one iteration function is constructed. In the fourth section, some properties and theorems are discussed. And we conclude the paper in the last section.

II. PRELIMINARY

As a new privacy-preserving model, differential privacy attracts many researchers' eyes. It is almost the most popular one. In this section, we will introduce some basic concepts and rules which we will use.

A. Notations

There are some notations used in the paper. In order to understand the paper easier, you can find their meanings in detail as following:

D : *dataset, and* D、D' *describe two neighboring dataset*;

M : *privacy mechanism*;

R : *output set*;

ε : *privacy budget*;

$Lap()$: *Laplace distribution function*;

$E()$: *the function to produce* ε;

$f()$: *the query function*;

Δf : *the sensitivity of query function* f;

i : *query count*;

B. Differential Privacy

Definition 1(Differential Privacy) *[8]* A randomized algorithm M gives ε - differential privacy for every set of outputs R , and for any neighboring datasets of D and D' , if M satisfies:

$$\frac{\Pr[M(D) \in R]}{\Pr[M(D') \in R]} \le e^{\varepsilon} .$$

In definition 1, parameter ε is called the ***privacy budget***. The privacy guarantee level of mechanism M could be deduced by ε . If ε is smaller, it shows that privacy preserving degree are higher. In practice, ε is usually set as less than 1, such as 0.1 or ln2.

Definition 2(Laplace Mechanism) *[8]* There is a dataset D and a function $f : D \to R$ over it, if the following condition equation is correct,

$$M(D) = f(D) + Lap(\frac{\Delta f}{\varepsilon})$$

Then mechanism M provides the ε -differential privacy .

In this formula, $Lap(\frac{\Delta f}{\varepsilon})$ stands for the size of the noise. So we can say, if *the sensitivity of query function* f (written as Δf) is larger and privacy budget ε is smaller, then the size of noise is

Actually, Laplace mechanism given by definition 2 can satisfy the attribute given in definition 1 through poof. That is theorem 1 as following:

Theorem 1 *[11]* The Laplace mechanism preserves ε -differential privacy.

C. Interactive and Non-interactive Framework

In this publishing scenario, there are two different frameworks: interactive and non-interactive framework. In the former framework, a query f_i cannot be issued until the answer to the previous query f_{i-1} has been published. Figure 1 can show you the details.

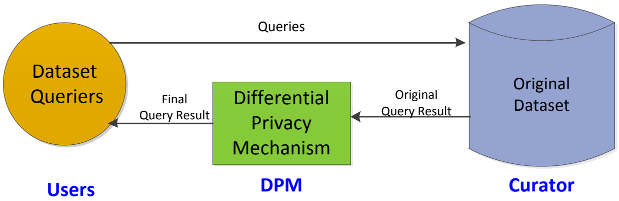

Figure 1. Interactive Publishing Framework

In the later framework, all queries are given to the curator at one time. The curator can provide answers with full knowledge of the query set. You can find the process in detail in figure 2.

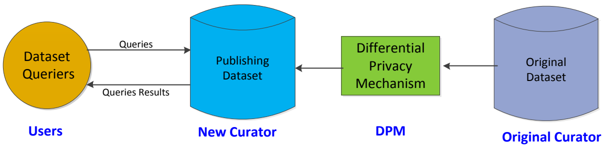

Figure 2. Non-interactive Publishing Framework

III. ADAPTIVE DIFFERENTIAL PRIVACY MODEL BASED ON DYNAMIC FEED-BACK (ADPM-DF)

A. Main Idea of the Model

The model deals with the problem of the privacy leakage which will increase with the query numbers enlarging in the interactive publishing scenario. We utilize dynamic feedback scheme to update the privacy budget. With the increase of query numbers, the privacy budget will decrease gradually. And the new privacy budget will do effect on the Laplace mechanism to the original query results.

According to the differential privacy mechanism, we can know that with increase of privacy budget, the utility of the mechanism will decrease gradually. The curator will not forbid the users' queries unfriendly, however, under this mechanism. At the same time, the privacy will not be leaked more with the query count enlarging. As a matter of fact, the users will "automatically" stop the query after some queries at last because of the better and better privacy preserving and worse and worse data utility.

B. Adaptive DP Model with Dynamic Feedback(ADPM-DF)

Based on the main idea given by the above, one adaptive differential model with dynamic feedback will be proposed as figure 3.

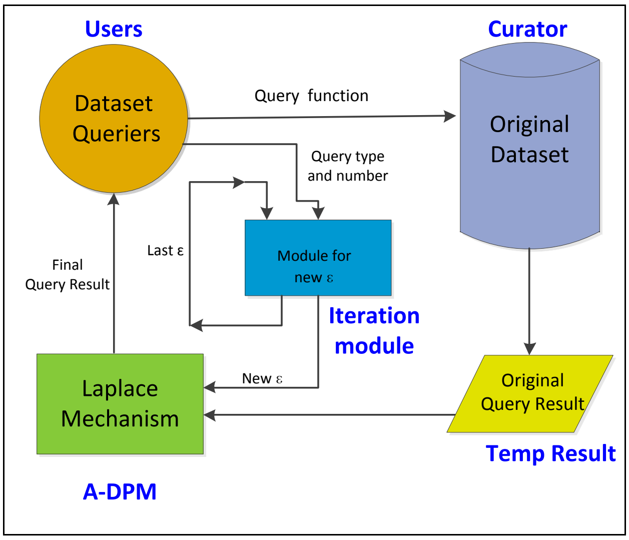

Figure 3. Adaptive Differential Privacy Model with Dynamic Feedback(ADPM-DF)

In this model, there are two different stages, one is the update of the differential privacy budget parameter ε. And in the other stage, the Laplace mechanism is used on the original query result with the new privacy budget parameter ε. The more important stage is the former one because it is the base of the second stage. So, we will discuss the feedback scheme to update privacy budget parameter ε in the next section.

C. Dynamic Feedback Scheme

In the model above, we adopt the feedback scheme. Actually, the feedback scheme is realized by one iteration function $E(\ )$ which can produce a new $\varepsilon_i (i \in Z^+)$ through the former parameter $\varepsilon_{i-1}(i \in Z^+)$ and the query count i. All the $\varepsilon_i (i \in Z^+)$ generated are nonnegative. And with the increase of i, $\varepsilon_i (i \in Z)$ will decrease gradually. So the function $E(\varepsilon)$ is monotone decreasing.

$$\varepsilon_i = E(\varepsilon_{i-1}, i) \quad , \ i \in Z^+;$$
$$s.t. \begin{cases} \varepsilon_0 \ \ is\ given \\ \varepsilon_i \geq 0 \\ \varepsilon_i \leq \varepsilon_{i-1} \end{cases} .$$

According the rule given by the iteration above, we can construct one example function. For example,

$$\varepsilon_{i+1} = (\frac{1}{2})^{\varepsilon_i + i};$$

$$\varepsilon_i = i;$$

$$i \in Z^+.$$

D. Technological Process of ADPM-DF

If the *ADPM-DF* will be used in practice, its technological process will be described as following Figure 4.

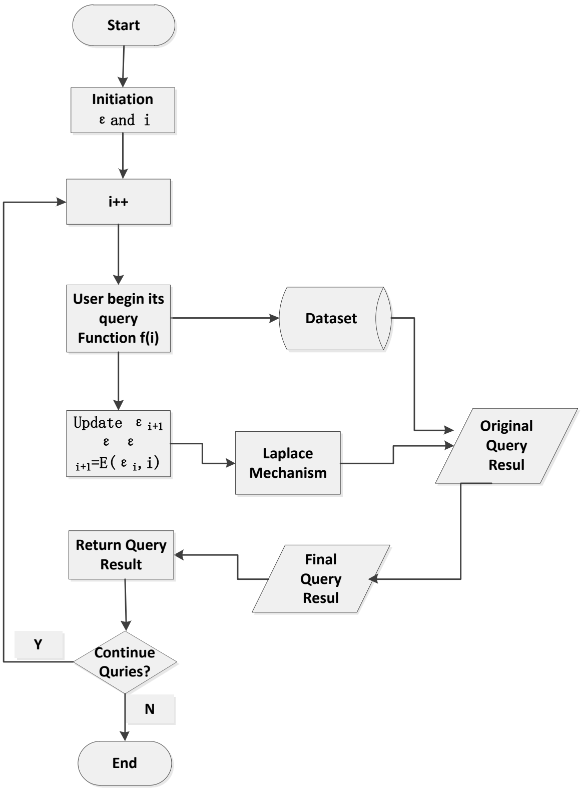

Figure 4. Technological Process of ADPM-DF

IV. MODEL ANALYSIS

In the model of adaptive DP based on dynamic feedback, there are several good qualities. These qualities will be analyzed and some theorems with their proofs are given in this section.

Firstly, we can find that the iteration scheme is monotonous decrease. With the increase of i, ε_i will decrease gradually. It is decided by the iteration function $E()$.

Based on this, the parameter ε_i in the feedback scheme can promise the differential privacy quality. It's given by theorem 4 as following.

Theorem 4 In the model of *ADPM-DF*, if the initiation of parameter ε_0 satisfied differential privacy, then all the new $\varepsilon_i(i \in Z^+)$ generated by the iteration algorithm above can satisfy differential privacy properties.

The proof of this theorem is omitted here. And through the theorem, inference 1 is very clear as following:

Inference 1 In the model of *ADPM-DF,* the degree of the privacy which the differential privacy mechanism could promise , will increase gradually.

V. *CONCLUSIONS*

In recent years, data publishing is very meaningful and necessary to the social development. However, most of the datasets to be published are full of personal information. Privacy preserving has become a serious challenge.

This paper adopts the popular privacy preserving method- differential privacy to deal with the privacy leakage with the increase of query count. The publishing framework we consider here is interactive publishing. In this paper, one new model *ADPM-DF* (adaptive differential privacy interactive publishing model based on dynamic feedback) is proposed. The feedback scheme is designed with an iteration algorithm to generate new privacy budget parameter. Some properties are discussed in the new model. Analysis shows that the new model can run well.

However，in this paper what we focus on is only the interactive publishing framework, which is one of two common publishing scenarios, including interactive and non-interactive publishing scenarios. And, during differential privacy preserving mechanisms, we select the Laplace mechanism which is suit for numerical data. So, how to deal with data privacy leakage in many times queries in non-interactive publishing scenario and how to deal with non-numerical data will be our work in the future.

ACKNOWLEDGEMENT

This work was partly supported by the National Natural Science Foundation of China under Grant 61402275, open project fund of Guizhou Provincial Key Laboratory of Public Big Data (2017BDKFJJ026), Shaanxi Province Natural Science Basic Research Program Funded Project (2016JM6069), the Fundamental Research Funds for the Central Universities under Grant (GK201402004, GK2018030005), the Scientific Research Foundation for the Returned Overseas Chinese Scholars of MOHRSS.

REFERENCE

[1] Weitzner D.J. Bruce E J. Big data privacy workshop: Advancing the state of the art in technology and practice. http://web.mit.edu/ bigdata-priv/index.html. 2014.

[2] T. Zhu et al., Differential privacy and applications, Advances in Information Security 69, DOI 10.1007/978-3-319-62004-6 , Springer International Publishing , AG 2017.

[3] Sweeney L. K-anonymity: a model for protecting privacy[J],international Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(5):557-570.

[4] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 1-52.

[5] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity[C].2007 IEEE 23rd International Conference on Data Engineering, 2007: 106-115.

[6] Xiao X, Tao Y. m-invariance: Towards privacy preserving re-publication of dynamic datasets[C].Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007:689-700.

[7] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24–28, 2016, pages 308–318, 2016.

[8] C. Dwork, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C].Theory of Cryptography Conference, 2006: 265-284.

[9] C. Dwork, K. Kenthapadi, F.McSherry, I.Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), pages 486–503, 2006.

[10] I. Dinur and K. Nissim. Revealing information while preserving privacy. In PODS,pages 202–210, 2003.

[11] C. Dwork,, Rothblum G N. Concentrated differential privacy. arXiv preprint arXiv:1603.01887,2016.

[12] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In STOC, pages 117–126, 2015.

[13] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In NIPS, pages 2350–2358, 2015.

[14] McSherry F, Talwar K. Mechanism design via differential privacy[C]. FOCS'07, 48th Annual IEEE Symposium on, 2007: 94-103.

[15] Q Geng, P Viswanath. The optimal noise-adding mechanism in differential privacy[J]. IEEE Transactions on Information Theory, 2016, 62(2): 925-969.

[16] Q Geng ， P Kairouz ， S Oh ， P Viswanath . The staircase mechanism in differential privacy. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(7):1176-1184.

[17] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science,2014,9(3-4):211−407.

[18] N Holohan ， DJ Leith ， O Mason . Optimal differentially private mechanisms for randomised response. IEEE Transactions on Information Forensics & Security, 2017, 12(11): 2726 – 2735.

[19] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science,2014,9(3-4):211−407.