A Workload Division Differential Privacy Algorithm to Improve the Accuracy for Linear Computations

Jun Li[1,2], Huan Ma[3], Guangjun Wu*[1](✉), Yanqin Zhang[4], Bingnan Ma[3], Zhen Hui[3], Lei Zhang[1], and Bingqing Zhu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{lijun,wuguangjun,zhubingqing,zhanglei}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] National Computer Network Emergency Response Technical Team/Coordination, Center of China (CNCERT/CC), Beijing, China
{mahuang, huizhen}@cert.org.cn, mabingnan90@gmail.com
[4] China Petroleum Engineering & Construction CORP. Beijing Design Company, Beijing, China
zhangyanqin.cpe@cnpc.com.cn

Abstract. Differential privacy algorithm is an effective technology to protect data privacy, and there are many pieces of research about differential privacy and some practical applications from the Internet companies, such as Apple and Google, etc. By differential privacy technology, the data organizations can allow external data scientists to explore their sensitive datasets, and the data owners can be ensured provable privacy guarantees meanwhile. It is inevitable that the query results that will cause the error, as a consequence that the differential privacy algorithm would disturb the data, and some differential privacy algorithms are aimed to reduce the introduced noise. However, those algorithms just adopt to the simple or relative uniform data, when the data distribution is complex, some algorithms will lose efficiency. In this paper, we propose a new simple ε-differential privacy algorithm. Our approach includes two key points: Firstly, we used Laplace-based noise to disturb answer to reduce the error of the linear computation queries under intensive data items by workload-aware noise; Secondly, we propose an optimized workload division method. We divide the queries recursively to reduce the added noise, which can reduce computation error when there exists query hot spot in the workload. We conduct extensive evaluation over six real-world datasets to examine the performance of our approach. The experimental results show that our approach can reduce nearly 40% computation error for linear computation when compared with MWEM, DAWA, and Identity. Meanwhile, our approach can achieve better response time to answer the query cases compared with the start-of-the-art algorithms.

* Guangjun Wu is the corresponding author of this paper.
This work was supported by the National Natural Science Foundation of China(No.61931019).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

Keywords: Differential privacy · Privacy-Preserving · Data Security.

1 Introduction

Among the data privacy protection technologies, existing research is based on the solution from the following perspectives: anonymity-based methods, encryption-based methods, noise-based method, and differential privacy-based method. There have been many reliable encryption-based method technology, such as DES[7], 3DES, Blowfish[23], RC5[21], IDEA, RSA, etc. The advance of the encryption technology is their security. However, the analyzability will be lost due to the encryption. The anonymity-based methods to protect data privacy can keep the data's analyzability, the mainly anonymity-based technologies are k-anonymity[24], L-diversity[3] and T-closeness[18]. However anonymity-based methods have fatal weaknesses, and the anonymous data might suffer anti-anonymity. For the data organizers, there exist security and privacy problems on data collection and publishing. Among the data privacy attacks, differential attack is a way that the attacker infers private data through statistical information over two homogeneous datasets. For example, an attacker can infer a person's specific shopping goods by differential attacks via different queries. To explore whether a person bought an *object*, the attacker can conduct two queries, and one query obtains the count of persons that have bought the *object*, and another query the count on the data set that excludes the person by the quasi-identifier, such as timestamp, gender, region, age, etc. By the two query results, the attack can infer whether the person bought the *object*.

To solve the differential attack, many differential privacy algorithms can be used, such as matrix mechanism[17], DAWA algorithm[16], MWEM[13], and RAPPOR[10], etc. The differential privacy technology can be used in many fields[26, 6, 5, 11, 20]. Differential privacy was first defined by Dwork et al[8, 9], and it protects the individual data by injecting noise to the results according to the privacy budget. A number of ε-differential privacy algorithms have been proposed[16, 13, 17, 2, 15], and some of them workload-aware and data-dependent[16, 13, 17, 2]. From the method of disturbing results view, ε-differential adopts three ways: Laplace Mechanism[8], Exponential Mechanism [19], and Randomized Response[25]. Random response mechanism is an effective way to protect the privacy of the frequency vector. The random response mechanism has been used in privacy protection of collecting sensitive data since the 1960s. RAPPOR[10] is ε-differential privacy technology that Google company has already used in the browser, and it adopts the random response. MWEM[13] is classical ε-differential privacy, and it is based on a combination of the Mechanism Exponential Mechanism with the Multiplicative Weights update rule. The MWEM algorithm selects and poses queries using the Exponential and Laplace Mechanisms, and it improves the approximation using the Multiplicative Weights update rule. DAWA[16] is a data-dependent and workload-aware algorithm, and it adds noise according to the input data and the given workload and it is a two-stage mechanism for answering range queries under ε-differential privacy. In 2016, Michael

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

Hay et al. propose an evaluation framework for standardized evaluation of privacy, called DPBENCH[14]. In 2018, Dan Zhang et al.[27] propose a programming framework and system called ϵktelo to implement the existing algorithms. For the task of answering linear counting queries, ϵktelo allows both privacy novices and experts to easily design algorithms, and the APEx[12]is a novel differential privacy system that allows data analysts to pose adaptively chosen sequences of queries along with required accuracy bounds.

Most of the algorithms are related to data distribution, especially when the data items are sparse, i.e., there are a large number of items are empty, these algorithms can effectively reduce the introduced errors. The same conclusion can be reached in the paper[16, 14]. While, these algorithms are not suitable for all data situations, as in the situation the data items are intensive and the data has complex distribution, and the conclusion is also shown in[16, 14]. Current ε-differential privacy algorithms will cause computation error for linear computations over the intensive data domain. Inspired by the partition of the data domain, we propose a novel ε-differential privacy algorithm via Laplace-based noise and optimized workload division to decrease the computation error in complex data situation. We make the following contributions:

(1) We propose a novel ε-differential privacy algorithm in complex data situation. We used Laplace-based noise to disturb the query results. This disturbation can reduce the error of the linear computation queries under intensive data items by workload-aware noise.

(2) We propose an optimized workload division method. We divide the queries recursively to reduce the added noise. This division can effectively reduce computation error when there exists a hot spot, i.e., some domain is frequently queried in the workload.

(3) We conduct extensive experiments on six real-world datasets and conduct a comparison with differential privacy algorithms (MWEM, DAWA, and Identity). The evaluation results show that the proposed algorithm can effectively reduce the computation error and has better efficiency relatively.

2 Approach Overview

We propose a ε-differential privacy algorithm for the linear computation queries. The algorithm aims to reduce the results error in the case that the sensitivity of workload is high and there exists frequency queried dom($\mathbb{B}$) item due to the hot issue or statistical attack queries and the frequency count x is complex. In the algorithm, we adopt Laplace Mechanism to disturb the query results. To reduce the random added noise, we propose a novel perspective that the added noise might be reduced by dividing the queries into several clusters and add Laplace-base noise respectively. Furthermore, based on the Laplace division, we propose a simple and effective recursion division for the query workloads and the privacy budget. The method recursively divides the queries workload and privacy budget into two parts when the expected noise is less than that before dividing. To sum up, the algorithm can solve three problems:(1)The current ε-algorithms can

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

reduce the error limitly, meanwhile, those algorithms will cost much computation resources. (2)When the sensitivity of a query workload is large, the current algorithms can't reduce the noise obviously. This can be shown in[16]. (3)In the situation that the data distribution is intensive, the current algorithms cannot fit it and will cause much error for the answer to query workload.

The method we propose satisfies ε-differential privacy rigorously, and ε-differential privacy is the privacy protection mechanism proposed by Dwork in 2006 and regulates privacy protection. We will define ε-differential privacy formally.

Definition 1 (ε-differential privacy). *An algorithm M is a ε-differential privacy algorithm if for any neighboring database I and I' ($|I - I'| \leq 1$), and any subset of output S satisfies the following formula:*

$$Pr[M(I) \in S] \leq \exp(\varepsilon) \times Pr[M(I') \in S]$$

The ε-differential privacy algorithm protects privacy data by disturbing the answer and the attackers cannot distinguish the results over the neighboring database I and I', and the parameter ε is the privacy budget and it determines the privacy-preserving capacity. If the privacy budget is lower, the differential algorithm will protect privacy more effectively. For the random algorithm M, if the results over the two adjacent datasets I and I' are close to each other, and it is difficult to infer whether a data item exists by $M(X)$ and $M(Y)$. ε-differential privacy has the following three primary properties.

Property 1. For the random algorithm ε_i-difference privacy M_1, and function $M(X)$ is an arbitrary deterministic function: $R \to R'$. Then $M_1(M(X))$ still satisfies ε differential privacy.

Property 2. For the random algorithm M_i and it satisfies ε_i-difference privacy. Defining a random function M that it is a process of a random sequence of M_i. The random function M satisfies $\sum_{i=1}^{k} \varepsilon_i$-difference privacy.

Property 3. The data set X make up of k data sets $\{X_1, ...Xi, ...X_k\}$, and $M_i(X_i)$ satisfies ε-differential privacy, respectively. $M(X) = \{M_1(X_1), ..., X_k(M_k)\}$ satisfies $\max_{\varepsilon_i} \varepsilon_i$-differential privacy.

Our algorithm reduces the results error when answering the linear computation query under the ε-differential privacy, and we will define the linear computation query. For a database instance I whose relational schema attributes $\mathbb{A} = \{A_1, A_2, ..., A_l\}$. In $\mathbb{A}$, each attribute data can be discrete or continuous. For the continuous data, the data can be treated as discrete in the data domain as well. The *workload* means a set of queries over the attributes $\mathbb{B} = \{B_1, B_2, ..., B_k\}, \mathbb{B} \in \mathbb{A}$. For example, if the *workload* queries in a subset of three-dimensional range query over attributes A_1, A_2, and A_3, $\mathbb{B} = \{A_1, A_2, A_3\}$. We then present a frequency vector x, and $x_i \in dom(\mathbb{B})$. For example, $\text{dom}(\mathbb{B}) = \{(1,1,1), (1,1,2), ...\}$ and for each $\text{dom}(\mathbb{B})_i$, x_i is the frequency of tuples values $\text{dom}(\mathbb{B})_i$. A linear computation query computes a linear combination of the frequency in x, as described

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

$$W = \begin{bmatrix} 1\,1\,0\,0\,0\,0 \\ 1\,1\,1\,0\,0\,0 \\ 0\,1\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 1\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix}, x_t = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \\ 0 \\ 9 \end{bmatrix}, S = W \cdot x^t = \begin{bmatrix} 5 \\ 10 \\ 7 \\ 3 \\ 4 \\ 1 \\ 0 \\ 9 \end{bmatrix} \quad (1)$$

Fig. 1: A sample of linear computation query workload, frequency vector, and answer to the workload.

the following SQL query and we define the linear computation query as follows definition formally.

$$Select\ count(*)\ from\ R\ Where\ dom(\mathbb{B}) = dom(\mathbb{B})_i\ or...\ dom(\mathbb{B}) = dom(\mathbb{B})_k$$

Definition 2 (Linear computation query). *A linear computation query is a length-n vector* $q = [q_1, q_2, ..., q_n]$ *, each* $q_i \in \{0,1\}$*. The answer to a linear query* q *on* x *is the vector product* $q \cdot x = q_1x_1 + q_2x_2 + ..., +q_nx_n$

The linear computation can be called range count query, linear count query, and point count query when the query q can be marked as range, length-n vector, or a position in x.

In the data collection situation, calculating the frequency in x can be done by the data organizers. And the data organizers has the capability to answer the linear computation query over the frequency vector x. The workload W makes up of a set of linear computation queries. If W is an $m \times n$ matrix, it means m length-n linear computation queries and the query results can be computed as the matrix product $W \cdot x$. The linear computation query is one of the most important and common queries in data mining and data analysis. The linear computation can help the analyst understand the distribution information of data and to make intelligent decisions and data prediction. Figure 1 shows a workload W, frequency vector x, and the answer to W over x.

3 Laplace-based Disturbation

Our algorithm adopts Laplace Mechanism to add noise, and we transform the Laplace Machainsim[8] to fit the query workload and data distribution. To ensure our algorithm satisfy ε-differential privacy, the algorithm adds random noise rigorously conform to the Laplace distribution.

3.1 Laplace Mechanism

The Laplace mechanism [8] is proposed by Dwork, and the key method of Laplace Mechanism is to add noise that randomly generated through the Laplace distribution to the query results. The probability density function of the Laplace

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

distribution is described as following:

$$Lab\left(x;a,b\right)=\frac{1}{2b}\exp\left(\frac{|x-a|}{b}\right) \tag{2}$$

The variance of a random variable that satisfies the Laplace distribution is $\sigma^2=2b^2$. To make the algorithm satisfy ε-differential privacy, we can add random noise from the $Lap(x;a,0)$, and we denote the Laplace distribution random variable as $Lap(a)$ in the following section. For different query or query workload, the Laplace Mechanism adds noise differs against the sensitivity of the query or workload.

Definition 3 (Sensitivity). *Given a query q and the frequency vector x and x', the sensitivity of the query q is:*

$$\Delta_q=\max\left|\left|q\left(x\right)-q\left(x'\right)\right|\right|_1 \quad \left(\left|\left|x-x'\right|\right|_1\leq 1\right)$$

It can be seen that the sensitivity of a query is the maximum change of the answer to a query on the neighboring frequency vectors. When the sensitivity of a query is high, the privacy data has a high probability to be attacked, and the reason is that the presence or absence of certain data can greatly change the result of the query, and it is more calculable to infer the certain sensitive data. For a query workload W, we use an $m\times n$ matrix to represent W, as shown in Figure 2. According to the sensitivity of a query, the sensitivity of the query workload W can be defined as the following:

$$\Delta_W=max\parallel Wx^t-Wx'^t\parallel=\max_j\parallel\sum_{i=1}^{i=m}\left|W_{ij}\right|,(\parallel x-x'\parallel_1\leq 1)$$

Given the definition of Laplace distribution and sensitivity, we can define the Laplace mechanism as following formally.

Definition 4 (Laplace Mechanism). *Given a workload W and a frequency vector x, $M_L(x,W,\varepsilon)$ is ε-differential privacy, if it satisfies the following condition:*

$$M_L(x,W,\varepsilon)=W\cdot x^t+(Y_1,\ldots,Y_k))$$

The random variable Y_i is generated by $Lap\left(\Delta_W\cdot\frac{1}{\varepsilon}\right)$. The proof is presented as the following, where database I and I differ at most one record, $P_I(s)$ is the probability that the output for the query database I is s.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

$$\begin{aligned}\frac{p_I(s)}{p_{I'}(s)} &= \prod_{i=1}^{k}\left(\frac{\exp\left(-\frac{\varepsilon\ \left|q(I)_i - s_i\right|}{\nabla q}\right)}{\exp\left(-\frac{\varepsilon\left|q(I')_i - z_i\right|}{\nabla q}\right)}\right) \\ &= \prod_{i=1}^{k}\left(\frac{\exp\left(-\frac{\varepsilon\ \left|q(I)_i - s_i\right|}{\nabla q}\right)}{\exp\left(-\frac{\varepsilon\ \left|q(I')_i - s_i\right|}{\nabla q}\right)}\right) \\ &= \prod_{i=1}^{k}\left(\exp\left(\frac{\varepsilon\left(\left|q\left(I'\right)_i - s_i\right| - \left|f\left(I\right)_i - s_i\right|\right)}{\nabla q}\right)\right) \\ &\leq \prod_{i=1}^{k}\left(\exp\left(\frac{\varepsilon\left(\left|q\left(I'\right)_i - q\left(i\right)_i\right|\right)}{\nabla q}\right)\right) \\ &\leq \exp\left(\varepsilon\ \right)\end{aligned} \tag{3}$$

3.2 Workload-aware Noise

To reduce the noise, we will divide the queries in workload into several workloads. Meanwhile, the privacy budget ε will be divided into the same number of privacy budgets. After dividing, different workloads will add corresponding noise according to the divided privacy budget. Formally, for the workload W, we divide it as $\{W_1, W_2, ..., W_m\}$, For each divided workload W_i, the privacy budget is also divided into $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_m\}$,and add random noise from the distribution $Lap(\Delta_{W_i}1/\varepsilon_i)$. That is, the answer to the workload is $S = [W_1, W_2, ..., W_m]\cdot x_t + [Lap(\Delta_{W_1}1/\varepsilon_1),\ Lap(\Delta_{W_2}1/\varepsilon_2), ...\ Lap(\Delta_{W_m}1/\varepsilon_m)]$. It can be proved that the algorithm satisfies ε-differential privacy as the *property*2, and we can also prove it by the following process. For the neighboring database I and I', that is, $\parallel I - I' \parallel_1 \leq 1$. Let $p_I(s)$ represent the distribution probability of query x on W, and $s \in R^k$:

$$\begin{aligned}\frac{p_I(s)}{p_{I'}(s)} &= \prod_{i=1}^{k}\left(\frac{\exp\left(-\frac{\varepsilon_i\left|q(I)_i - s_i\right|}{\nabla q_i}\right)}{\exp\left(-\frac{\varepsilon_i\left|q(I')_i - z_i\right|}{\nabla q_i}\right)}\right) \\ &= \prod_{j=1}^{m}\prod_{i\in W_j}\left(\frac{\exp\left(-\frac{\varepsilon_i\left|q(I)_i - s_i\right|}{\nabla q_i}\right)}{\exp\left(-\frac{\varepsilon_i\left|q(I')_i - s_i\right|}{\nabla q_i}\right)}\right) \\ &\leq \prod_{j=1}^{m}\exp\left(\varepsilon_j\right) \\ &= \exp\left(\varepsilon\right)\end{aligned} \tag{4}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

We discuss the error change by the dividing for workload $W = \{W_1, W_2, ..., W_m\}$, and privacy budget $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_m\}$. We calculate the average L_1 error for the answer to the workload. The excepted L_1 error of the answer before dividing and after dividing the workload is as follows:

$$E\left(\mid Lap(\Delta_W \cdot 1/\varepsilon) \mid\right) = \Delta_{W_i} \cdot \frac{1}{\varepsilon_i}$$

$$E\left(\frac{1}{k}\sum_{i=1}^{m} \mid Lap(\Delta_{W_i} \cdot 1/\varepsilon_i) \mid \cdot \mid W_i \mid\right) = \frac{1}{k}\sum_{i=1}^{m} \Delta_{W_i} \cdot \frac{1}{\varepsilon_i} \cdot \mid W_i \mid$$

4 Optimized Workload Division

Taking the above workload in Figure 2 for example, the original workload and divided workloads as following. When the workload W adopts 1-differential privacy by the Laplace Mechanism. The excepted L_1 error is $\Delta_W/\varepsilon = 4$, and after

$$W = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&1&1&0&0&0 \\ 0&1&1&0&0&0 \\ 0&0&1&0&0&0 \\ 0&0&1&0&0&0 \\ 0&0&0&1&0&0 \\ 1&1&0&0&0&0 \\ 0&0&0&0&0&1 \end{bmatrix}, W_1 = \begin{bmatrix} 1&1&0&0&0&0 \\ 1&1&1&0&0&0 \\ 0&1&1&0&0&0 \end{bmatrix}, W_2 = \begin{bmatrix} 0&0&1&0&0&0 \\ 0&0&1&0&0&0 \\ 0&0&0&1&0&0 \\ 1&1&0&0&0&0 \\ 0&0&0&0&0&1 \end{bmatrix} \tag{5}$$

Fig. 2: A sample of dividing for workload.

dividing the workload into W_1 and W_2, the privacy budget into $\varepsilon_1 = 0.58$ and $\varepsilon_2 = 0.42$, the L_1 error will be 3.4275.

Basing on the dividing for workload and privacy budget, we propose a specific division in the data situation that the data is relatively large and the distribution of data is complex. We take the mean square error of the frequency vector x to discriminate the data distribution complexity. And in query workload, there exist data domain queried with high frequency. To reduce the added Laplace-based noise, we divide the privacy budget into two equal parts iteratively, and the workload is divided according to the sensitivity, and the process can be described in Algorithm 1.

The dividing in the algorithm will continue until the recursion finished. We will discuss the rationality of the division. The dichotomy is used as the reason that the query workload and privacy budget are divided into two parts W_1, W_2, ε_1, ε_2, and for $E(L_1) = \frac{\nabla_{W_1}}{\varepsilon_1} * |W_1| + \frac{\nabla_{W_2}}{\varepsilon_2} * |W_2|$, $\varepsilon_1 = \varepsilon_2$ is the minimum extreme point of the function. As described in Algorithm 1, at first, we set the data domain queried by high frequency as high-frequency items. For workload

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

Algorithm 1 Workload dividing

procedure DIVIDEWORKLOAD(W, ε)
  $\varepsilon_1 = \varepsilon_2 = \varepsilon/2.0$
  get the most frequent item in x as x_i,
  $|W_1| = (|W| + \nabla_W + g)/4$
  select randomly $|W_1|$ queries as W_1 from W where x_i is queried and the rest queries as W_2
  $noise = |W| * \nabla_W$, $noise\_divided = |W_1| * \nabla_{W_1} + (|W| - |W_1|) \cdot \nabla_{W_2}$
  if $noise \geq noise\_divided$ **then** ▷ Stop dividing while noise doesn't reduce
    return ($DIVIDEWORKLOAD(W_1, \varepsilon/2)$,divideWorkload($W_1, \varepsilon/2$))
  return W ▷ return W while it is unnecessary to divide

W, its division is W_1 and W_2 and supposing that the sensitivity of W is the sum of W_1 and W_2. The total expected noise under the ε-differential privacy is

$$E(noise(\varepsilon_1, \varepsilon_2, W_1, W_2)) = \frac{\Delta_{W_1}}{\varepsilon_1} * |W_1| + \frac{\Delta_W - \Delta_{W_1}}{\varepsilon_2} * (|W| - |W_1|) \quad (6)$$

We can infer that $(\varepsilon/2, \varepsilon/2, W_1, W_2)$ is a point of minimum, so we adopt $\varepsilon_1 = \varepsilon 2 = 2/\varepsilon$. To get $min\ E(noise(\varepsilon_1, \varepsilon_2, W_1, W_2))$, we set $\Delta_{W_1} = |W_1|$, and we can compute that when $|W_1| = (W + \Delta_W)/4$, the $E(noise\varepsilon_1, \varepsilon_2, W_1, W2)$ will be a minimal value. The parameter g can optimize the result as a consequence of that for a workload W and its divisions W_1,W_2, $\Delta_W > (\Delta_{W_1} + \Delta_{W_2})$, which is not in accordance with our assumption. Therefore, we introduce the parameter to regulate the result and the g can be estimation by the specific workload.

5 Experimental Evaluation

We now evaluate the performance of our approach on multiple datasets and workloads and compare our algorithm with state-of-the-art differential privacy algorithms. The main metric is average error, and we evaluate the metric on differential datasets and workloads.

In follow evaluation, we test our algorithm with the metric average L_1 error per query result of the given workload. The workloads we use are generated randomly and the data set is from the real public database. To make the result more convincing,we run 5 trials for each evaluation. Furthermore, we test the time efficiency of our algorithm.The synthetic workload also is used by all the comparison algorithms. In the experiment, we set the privacy budget varies in {10.0,5.0,1.0,0.5,0.1}.In the following sections, we describe the datasets, workload, and the parameters in the experiment. In the above section, we have described the properties of liner count query. When there are multiple attributes in datasets, we can still use one-dimensional frequency vector x to represent the datasets. In the experiment, we use one-dimensional data sets. We use six real data sets. Adult comes from American statistical data[4]. The frequency vector x is built on the attribute "capital loss", which is also used in the experiment[13].

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

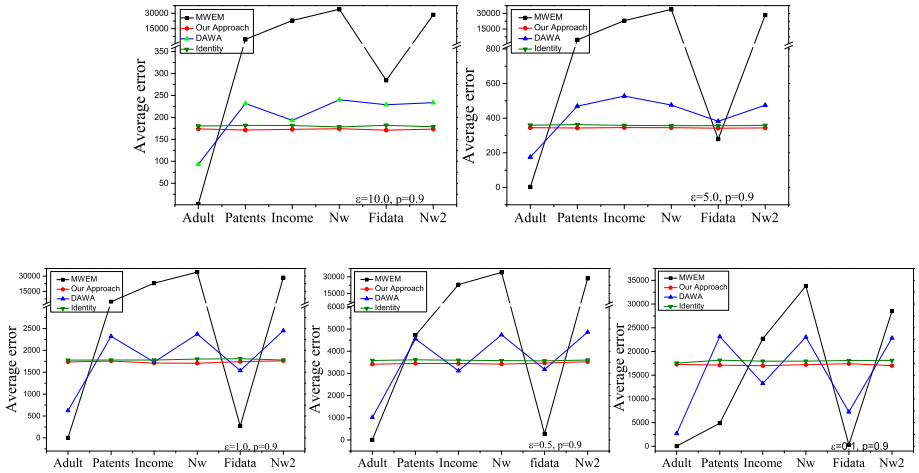

Fig. 3: Average error on the workload that the frequency count item is queried with probability as p=0.9

Table 1: Overview of the datasets in the experiments.

Datasets name	Scale	% Zero Count	Mean value	Variance
Adult	17665	97.998	4.31274	263.04404
Patents	27948226	6.20118	6823.29736	3532.42422
Income	20787122	44.971	5074.98095	47859.49063
Nw	32287151	0.268	7882.60522	60262.21603
Fidata	3519442	58.178	859.23880	18942.96715
Nw2	32678757	0.0	7978.21216	84866.75896

The Adult is sparse, and many frequency counts in x are zero. Income is from IPUMS American community survey data from 2001-2011, and frequency vector x is the count of personal Income[22], and Income is also used in DAWA[16]. Patent is a citation network among a subset of US patents[16]. Fidaeta is from census of fatal occupational injuries in the United States of American labor statistics[1], and both Nw and Nw2 are from a survey of compensation in the United States of American labor statistics[1], and they are the frequency vector by setting unit as 1 and 2 in the continuous value attribute. We take the length of x of the five datasets as 4096. The overview of datasets is described in Table 1.

For the query workload, we conduct the experiment on eight synthetic query workloads W. For the frequently queried item in frequency vector x, we set probability of being queried $p = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. A workload has 2000 queries, and each query $q \in W$ randomly selected a center cluster, and the frequency counts in x are randomly generated via the normal distribution with c

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

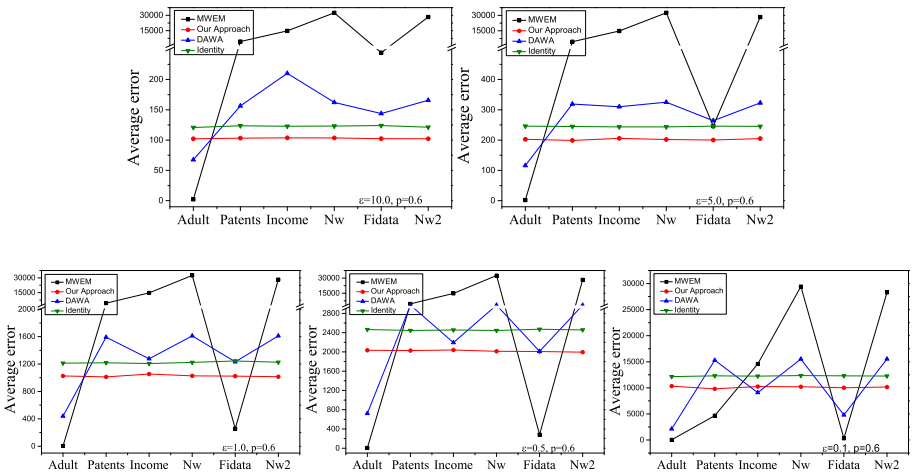

Fig. 4: Average error on the workload that the frequency count item is queried with probability as p=0.6

as the center and 10 as the variance. Furthermore, we compare three algorithms with our algorithm. The Identity[8] algorithm adopts Laplace Mechanism that the answer results are directly added Laplace distribution noise for disturbance. MWEM[13] achieves differential privacy technology by obtaining an estimate of x through Laplace Mechanism and Exponential Mechanism. DAWA [16] algorithm adopts the partitioning method to achieve the differential privacy for range count workload and linear count workload.

Among the experimental datasets, Adult is a "sparse" data set. The data distribution is relatively even-distributed as shown in Table 1, and zero accounts for more than 97% in the frequency vector x. The other four experimental data sets are "complex" data sets with a large scale and complex data distribution. Figure 4, 5, and 6 show the L_1 average error for the parameter p as 0,9, 0.6 and 0.2. It can be seen that MWEM[13] and DAWA[16] will add more noise than the Identity [8] algorithms, meanwhile, our algorithm always adds less noise than the Identity. The results figures show that MWEM[13] and DAWA[16] algorithm are datasets-aware and when facing different datasets, both algorithms perform differently over the same workload. The MWEM is most erratic, and when the data sets are simple or approximately even-distributed, the algorithm can add less noise than the other algorithms, but not for the complex data. In Figure 7, we compare the discount of L_1 average error by comparing it with the Identity[8]. In the experiment, we compare the different perforation with different parameter p, which represents frequency of a certain $dom(\mathbb{B})_i$ in x. Figure 5 shows that in the experiment sets, when $p = 0.2$, the algorithm can reduce more than 40% the L_1 error than the Identity.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

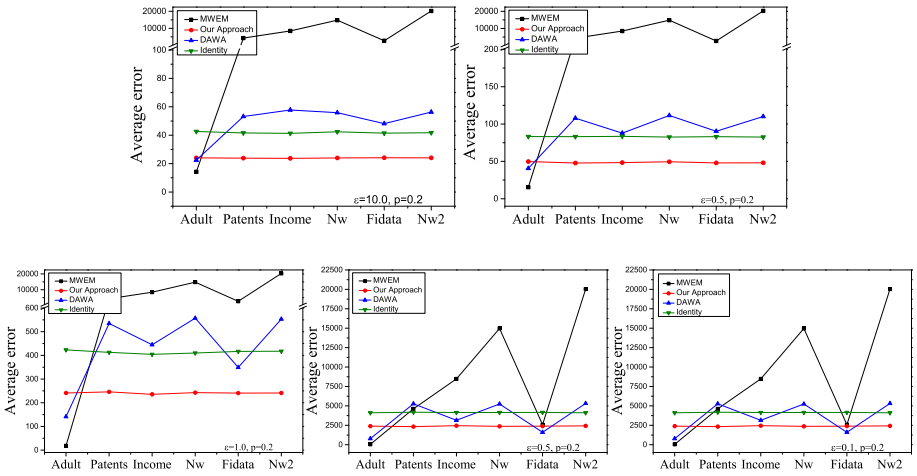

Fig. 5: Average error on the workload that the frequency count item is queried with probability as p=0.2

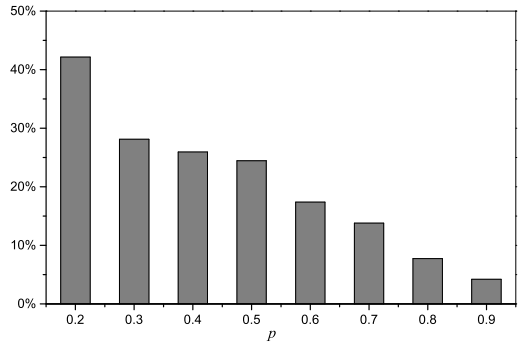

Fig. 6: The decrement of the average error by comparing our method with Identity.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

6 Conclusions

The ε-differential privacy is an effect privacy-preserving technology for linear computation. It can prompt data organizers to provide a secure third-party interface for statistical query. In this paper, we propose a novel ε-differential privacy algorithm, which uses Laplace-based noise and optimized workload division to decrease the computation error in complex data distribution for linear computations. The evaluation results show that our approach can reduce nearly 40% computation error when compared with the start-of-the-art differential privacy algorithms MWEM, DAWA, and Identity. As further work, we plan to extend our approach by optimizing the proposed work load division to reduce the introduced error.

References

1. https://www.bls.gov
2. Acs, G., Castelluccia, C., Chen, R.: Differentially private histogram publishing through lossy compression. In: 2012 IEEE 12th International Conference on Data Mining. pp. 1–10. IEEE (2012)
3. Ashwin, M., Daniel, K., Johannes, G., Muthuramakrishnan, V.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data **1**(1), 1–52 (2007)
4. Bache, K., Lichman, M.: Uci machine learning repository, 2013. URL http://archive. ics. uci. edu/ml **5** (2013)
5. Backes, M., Meiser, S.: Differentially private smart metering with battery recharging. In: Data Privacy Management and Autonomous Spontaneous Security, pp. 194–212. Springer (2013)
6. Blocki, J., Blum, A., Datta, A., Sheffet, O.: Differentially private data analysis of social networks via restricted sensitivity. In: Proceedings of the 4th conference on Innovations in Theoretical Computer Science. pp. 87–96. ACM (2013)
7. Diffie, W., Hellman, M.E.: Special feature exhaustive cryptanalysis of the nbs data encryption standard. Computer **10**(6), 74–84 (1977)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. pp. 265–284. Springer (2006)
9. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
10. Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. pp. 1054–1067. ACM (2014)
11. Feng, P., Zhu, H., Liu, Y., Chen, Y., Zheng, Q.: Differential privacy protection recommendation algorithm based on student learning behavior. In: 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE) (2018)
12. Ge, C., He, X., Ilyas, I.F., Machanavajjhala, A.: Apex: Accuracy-aware differentially private data exploration. In: Proceedings of the 2019 International Conference on Management of Data. pp. 177–194. ACM (2019)
13. Hardt, M., Ligett, K., McSherry, F.: A simple and practical algorithm for differentially private data release. In: Advances in Neural Information Processing Systems. pp. 2339–2347 (2012)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33

14. Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., Zhang, D.: Principled evaluation of differentially private algorithms using dpbench. In: Proceedings of the 2016 International Conference on Management of Data. pp. 139–154. ACM (2016)
15. Lee, J., Clifton, C.W.: Top-k frequent itemsets via differentially private fp-trees. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 931–940. ACM (2014)
16. Li, C., Hay, M., Miklau, G., Wang, Y.: A data-and workload-aware algorithm for range queries under differential privacy. Proceedings of the VLDB Endowment **7**(5), 341–352 (2014)
17. Li, C., Hay, M., Rastogi, V., Miklau, G., McGregor, A.: Optimizing linear counting queries under differential privacy. In: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 123–134. ACM (2010)
18. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. pp. 106–115. IEEE (2007)
19. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: FOCS. vol. 7, pp. 94–103 (2007)
20. Pham, A.T., Raich, R.: Differential privacy for positive and unlabeled learning with known class priors. In: 2018 IEEE Statistical Signal Processing Workshop (SSP) (2018)
21. Rivest, R.L.: The rc5 encryption algorithm. In: International Workshop on Fast Software Encryption. pp. 86–96. Springer (1994)
22. Ruggles, S., Alexander, J.T., Genadek, K., Goeken, R., Schroeder, M.B., Sobek, M.: Integrated public use microdata series: Version 5.0, 2010. Minnesota Population Center, Minneapolis, MN (2015)
23. Schneier, B.: Description of a new variable-length key, 64-bit block cipher (blowfish). In: International Workshop on Fast Software Encryption. pp. 191–204. Springer (1993)
24. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)
25. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association **60**(309), 63–69 (1965)
26. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., Winslett, M.: Differentially private histogram publication. The VLDB Journal—The International Journal on Very Large Data Bases **22**(6), 797–822 (2013)
27. Zhang, D., McKenna, R., Kotsogiannis, I., Hay, M., Machanavajjhala, A., Miklau, G.: Ektelo: A framework for defining differentially-private computations. In: Proceedings of the 2018 International Conference on Management of Data. pp. 115–130. ACM (2018)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_33