Difference Privacy Histogram Release Based on Isotonic Regression

Xiujin Shi*, Ling Zhou
School of Computer Science and Technology, Donghua University, ShangHai, China.

* Corresponding author. Tel.: 18671372186; email: zhouling189@gmail.com
Manuscript submitted August 10, 2015; accepted November 10, 2015.
doi: 10.17706/jsw.11.1.1-9

Abstract: Data release is likely to result in privacy disclosure, so appropriate privacy protection measures are required for various data release technologies in order to ensure the privacy and safety of information, while differential privacy as a reliable model for privacy protection is extensively researched and applied. This paper presents the histogram data publishing solutions under differential privacy model, namely adding noise on the optimized histogram structure and then carrying out isotonic regression algorithms on the histogram privacy sequence. In this case, differentcial privacy model keeps all the statistical properties of the histogram unchanged and the concealment of privacy information, and in addition, histogram reconstruction and isotonic regression algorithm are effective in improving the accuracy of data release via histogram. This paper provides a solution about isotonic regression to decrease the error on histogram reconstruction based on previous research.

Key words: Data release, differential privacy, histogram, isotonic regression.

1. Introduction

Along with higher social informatization, information released in any occasion can lead to disclosure of private information, hence the issue of privacy protection. The subjects of PPDR [1] (Privacy Preserving Data Release) researches all focused on accurate and effective information release for the purpose of making all users including legal ones and potential attackers inaccessible to accurate information of any individual when accessing data.

In previous studies, researchers pioneered the use of anonymity as the data release privacy-preserving mechanism [2], [3]. Though some researchers in 2002 found that private information contained in the data still could be obtained through certain attacks, this algorithm model is instructive to later researches on privacy protection. In view of the application and development of K Anonymity Algorithm, Dwork and his team in 2006 first proposed privacy-preserving method based on data distortion [4], [5], namely differential privacy model. The model is effective in protecting sensitive data from attacks of high probability inference, which fundamentally addresses the deficiencies in K Anonymity Algorithm.

Nowadays differential privacy algorithm has been a recent and widely applied model for privacy protection, and is significant for sensitive data release. Data released under differential privacy model in virtue is the approximate version of real data distribution, for they share almost the same statistic properties, and the information entity described by the data will not be disclosed at the time of data release [6].

Histogram is an important technique to effectively capture the centered distribution characteristic of the data as well as the data release means frequently used in PPDR researches through which privacy data release can be further analyzed and processed. In terms of the data release result, histogram, which

presents the query result of database aggregation or the marginal distribution of dataset, can reveal the distribution of the whole dataset in a specified dimension.

This paper carries out noise adding via differential privacy model, and achieves the congruence of the data sequence by isotonic regression in the process of noise adding and thus further improves the accuracy of privacy data release via histogram. Section Two is the introduction of differential privacy model and histogram. Section Three is histogram error analysis and histogram reconstruction through dynamic programming algorithm. Section Four is the discussion of the realization and application of isotonic algorithm for histogram data release. Section Five is experiments with transaction data and comparative analysis of experimental errors. Section Six is the conclusion of the research.

2. Related Work

2.1. Differential Privacy

The basic idea of differential privacy model is concealing the individual information in the data by means of data distortion based on the result of data processing to render attacks through high-probability inference impossible.

For example, when dataset D contains an information entity Alice, sensitive individual information can be disclosed on the release of $Q(D)$, the result of statistical measures including count, average and extreme value on related attributes including medical history and income level and other sensitive information in the dataset. Differential privacy model offers a good solution to such a problem: even if the information entity Alice in the dataset is deleted, the statistical result remains the same, which means that the disclosure risk of Alice's information, if any, will not increase for its involvement in statistical work. To sum up, differential privacy model ensures that algorithm outputs of two datasets with one piece of record numerically different are statistically approximate [7]. The standard definition of differential privacy is as follows.

Definition 1 (Differential Privacy)[8] Given two datasets $D1$ and $D2$ with identical structures, also called contiguous datasets, and dom(A) as the range of stochastic algorithm A with output $O \in \text{dom}(A)$, if

$$\Pr[A(D1) = O] \le e^{\varepsilon} \times \Pr[A(D2) = O] \tag{1}$$

then algorithm A satisfies ε-differential privacy, in which Pr[.] is the probability of stochastic output, and ε is differential privacy budget.

From Definition 1, it can be seen that differential privacy model clearly is distorting the effect of some data record on the algorithm output to make the output unaffected by the existence of the data record. Distortion is to make algorithm A in theory satisfy ε-differential privacy, and this paper employs Laplace Mechanism to achieve it.

Laplace Mechanism [8] is mainly applied in privacy-preserving occasions where the output is numeric value, such as linear query results and statistical results. Its realization lies in adding to the data stochastic noise following Laplace distribution whose parameter and privacy budget ε are correlated to algorithm sensitivity $\triangle$. The formula of such correlation is in which d is the dimension of database D.

$$A(\tilde{D}) = A(D) + Lap(\Delta_{\mathrm{A}} / \varepsilon)^{\mathrm{d}} \tag{2}$$

Definition 2 (Lp-sensitivity)[8] . Suppose that there is a function $f : D \to R^d$, in which D is the input dataset, Rd is the output range and $D1$ and $D2$ are contiguous databases, and p is the Lp_standard distance of the sensitivity of the measure f, $p \geqslant 1$, then the sensitivity of the function f is

$$\Delta_{\mathrm{p}}(f) = \max_{\mathrm{D1,D2}} \| f(D1) - f(D2) \|_p \tag{3}$$

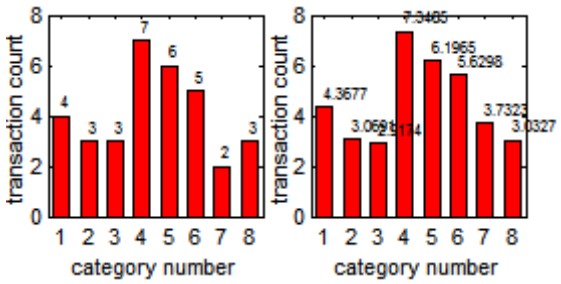
In specific application, $L1$ or $L2$ is usually used for calculation. In Laplace Mechanism, a algorithm with a smaller $\triangle$ will be used to reduce the consumption degree of privacy budget as much as possible.

2.2. Histogram Structure

The means of data release used in the paper is release via histogram which is an important tool to present data release. Through the histogram, data receivers can visually obtain data distribution in a specific dimension, and the released data can be used for further statistical analysis. The data release via histogram based on differential privacy model means specifically distributing privacy consumption in each interval according to the given privacy budget, and then adding noise to statistical measures in the histogram within the privacy consumption range to achieve data distortion.

Suppose that a dataset is divided into n intervals of single value or equal width in some dimension, hence the statistical sequence D={x_1, x_2, x_3, ……x_n}. Therefore, the basic form of differential privacy algorithm based on histogram release can be $\tilde{D}=D+\Gamma$, in which Γ is stochastic noise vector following Laplace distribution with 0 as the average value and $2/\varepsilon$ 2 as the variance, and $\Gamma=\{\tau_1,\tau_2.....\tau_n\}$, hence the statistic value of each interval in the histogram is $\tilde{x}=x+\tau$.

For example, suppose that the transaction records of selected items are compiled according to different categories in a transaction table, the statistical sequence of the original transaction records is D={4,3,3,7,6,5,2,3}, and its histogram is shown as Fig 1(a); the stochastic noise vector τ ={0.36, 0.0691, -0.0826, 0.3485, 0.1965, 0.6298, 1.7323, 0.0327} comes from Laplace distribution with parameter (0, 1), so the statistical sequence after noise injecting is $\tilde{D}$ ={4.3677,3.0691,2.9174,7.3485,6.1965, 5.6298,3.7323,3.0327}, and its histogram is shown as Fig. 1(b).



(a) Before (b) After

Fig. 1. Histogram before and after noise injecting.

As the histogram shown above, though the goal of individual privacy protection in some extent is achieved with noise injecting, error in histogram arises directly from the noise, and this error is called noise error.

3. Realization of Histogram

Data release via histogram based on differential privacy model is applying reasonable and effective privacy protection algorithm to original data sequence and improving accuracy of data release and reliability of its result within privacy budget. The realization process of differential privacy histogram can be divided into four steps: original sequence, reconstructed histogram, histogram with noise, isotonic histogram. The detailed procedure can be seen in Fig. 2 as follows.

This section will analyze the noise error and structure error of the differential privacy histogram, and propose dynamic programming algorithm for its structure error to reconstruct histogram.

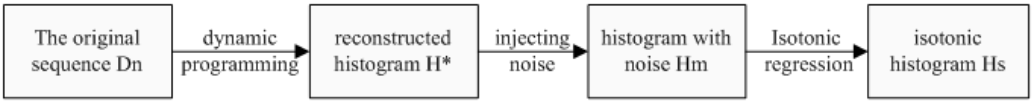


Fig. 2. Realization procedure of histogram.

3.1. Noise Error of Histogram

Given a database sequence $D=\{x_1, x_2, x_3, \cdots\cdots x_n\}$, H and H^* are the respective histogram structure of the sequence before noise adding and after, so the histogram error incurred by noise adding is

$$\text{Nerror}(H^*, H) = \sum_{i=1}^{n} (x_i - \tilde{x}_i)^2 = \sum_{i=1}^{n} \tau_i^2 \tag{4}$$

which in essence is the variance sum introduced by the noise of histogram, and is also the basic measure of noise error of the histogram. The expectation value of noise error of the histogram is

$$E(\text{Nerror}(H^*, H)) = E(\sum_{i=1}^{n} \tau_i^2) \tag{5}$$

If τ follows Laplace distribution with parameter $(0, \delta)$, then $E(\text{Nerror}(H^*, H))=n\delta^2$.

From the above measure of noise error of the histogram, it can be seen that the general expectation value of noise error of the histogram will increase in proportion to the parameter $|\delta|$ followed by the added noise. Therefore, under differential privacy requirement, reduction of the added noise parameter $|\delta|$ can effectively reduce the general noise error of the histogram.

If statistical process of intervals with equal width in each histogram is seen as aggregate query of conditions, according to the definition of algorithm sensitivity mentioned above, the query sensitivity of the histogram shown as Fig. 1 is $\triangle=1$, which means the maximum range change of histogram is 1 when any arbitrary transaction record in the database is deleted. According to the definition of Laplace, the variance followed by the noise added to the histogram is $2/\varepsilon^2$ in which ε is the given privacy budget. Therefore, with the privacy budget given, reduction of the query sensitivity $\triangle$ of the histogram can effectively reduce the noise error expectation of the histogram H, namely $E(\text{Nerror}(H^*, H))$, which makes the histogram structure more accurate based on privacy requirement.

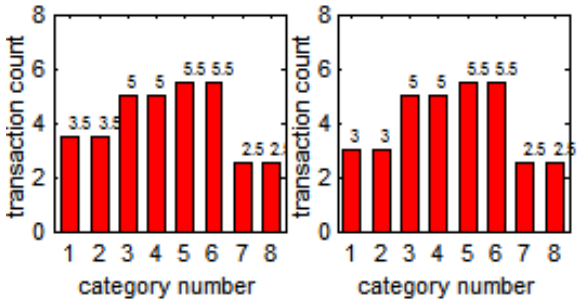
3.2. Reconstruction of Histogram and Its Error

From the above analysis of noise error of the histogram, it can be seen that reduction of the query sensitivity $\triangle$ of the histogram can improve the accuracy of data release. This section will argue about histogram reconstruction by dynamic programming algorithm, combining congruous intervals and employing the average value of the combined intervals as the new measure to replace the statistical result of the original interval.

Therefore, the query sensitivity of the histogram will change with the group number of combination. For example, the histogram in Fig. 1(a) combines every two congruous intervals of single value and replaces the statistical result of the original single value interval with the average value, hence the reconstruction sequence of the histogram D'=[3.5, 3.5, 5.0, 5.0, 5.5, 5.5, 2.5, 2.5], and the histogram is shown as Fig. 3(a).

As shown in the figure above, if any transaction record in Category 1 is deleted, then the statistical sequence of the congruous database is Dn={3, 3 , 3 , 7 , 6 , 5 , 2 , 3}. According to the reconstruction method shown in Fig. 3(a), the reconstruction sequence is Dn={3, 3 , 3 , 7 , 6 , 5 , 2 , 3}, and its histogram is shown as Fig. 3(b). In Fig. 3(b), the average statistical result of the transaction records of Commodity 1 and 2 is not 3.5 but 3 due to a transaction record of Commodity 1 is deleted, and thus the query sensitivity of the

histogram is 0.5 under this structure.



(a) before deletion (b) after deletion

Fig. 3. Histogram reconstruction.

From the example above, histogram reconstruction can effectively reduce query sensitivity and then noise error, and improve histogram release accuracy. The dynamic programming algorithm [9] is originally used for obtaining the optimal solution in decision process. It transforms the multistage process to a series of single-stage problems and uses intermediate results of all stages through algorithm to recursively obtain the optimal decision-making method.

In the process of histogram reconstruction, the intermediate result of dynamic programming algorithm $M(s, t)$, key to obtaining the optimal structure, is the minimum error when histogram sequences $x1$ to xs are reconstructed to t groups [10]. The matrix form of the intermediate results of the histogram based on dynamic programming in Fig. 1(a) is shown in Fig. 4, and because of the requirement of group reconstruction $s \geqslant t$, it is the upper part of the whole matrix form. The calculation formula of the intermediate results is $M(s,t) = \min_{t\text{-}1 \le k \le s\text{-}1} (M(k,t\text{-}1) + \text{Berror}(t))$, which specifically means to apply recursive method to t groups with error results of the previous t-1 groups in order as the arithmetic data, and then to select the group with the minimum structure error as the intermediate result of the t groups.

$$M(s,t) = \begin{bmatrix} 0 & 0.5 & 0.67 & 10.75 & 13.20 & 13.33 & 19.43 & 20.88 \\ & 0 & 0 & 0.67 & 1.17 & 2.67 & 13.33 & 13.83 \\ & & 0 & 0 & 0.5 & 1.17 & 2.67 & 3.17 \\ & & & 0 & 0 & 0.5 & 1.17 & 1.67 \\ & & & & 0 & 0 & 0.5 & 1 \\ & & & & & 0 & 0 & 0.5 \\ & & & & & & 0 & 0 \\ & & & & & & & 0 \end{bmatrix}$$

Fig. 4. Intermediate result $M(s, t)$.

Suppose the histogram with sequence length n is distributed to j groups, each group with i_j sequence values. If the original statistical sequence value is replaced with the average value of each group, and the bin error of j Group is

$$\text{Berror}(j) = \sum_{i_j} (x_{i_j} - \bar{x}_j)^2, \text{ and } \sum_j i_j = n \tag{6}$$

then the general structure error of the histogram is

$$\text{Serror} = \sum_j \text{Berror}(j) = \sum_j \sum_{i_j} (x_{i_j} - \bar{x}_j)^2 \tag{7}$$

Histogram reconstruction can effectively reduce the query sensitivity and improve the release accuracy, but if the structure error incurred by reconstruction process is relatively large, reduction of the noise error

of the histogram can be insignificant and thus unable to realize accurate release.

4. Realization of Isotonic Algorithm

Noise is added to meet the requirement of privacy protection in the process of sequence reconstruction, during which, however, each sequence value loses certain information, and the order constraint of the sequence is destroyed [11].

For example, in the original sequence Fig. 1(a) represents $x7<x8$, but after noise adding, in the sequence Fig. 3(b) represents $\tilde{x}7>\tilde{x}8$. Theretofore, under this circumstance, if the sequence after noise adding is regulated according to the order constraint of the original sequence, the privacy properties of the histogram will not be destroyed at all and, instead, its sequence accuracy will improve and so does the query accuracy. The regulation of sequence according to order constraint is realized by isotonic regression algorithm.

Often used in numerical value analysis, isotonic regression [12] uses the minimum data correction value to guarantee data order. For example, given a sequence s, require to determine sequence s' when minimum value of ||s-s'|| is achieved, and when $1 \leqslant i \leqslant n$, s'[$i$]$\leqslant$s'[$i$+1]. The determination of sequence s' starts with regulation of the first element of sequence s, then replaces the disordered subsequence tested with its average value, and then determines whether the average value and its next value is in order, and repeat this process until the last value [11]. The detailed algorithm procedure is shown as Algorithm 1.

Algorithm 1. Isotonic regression algorithm: input sequence s; outpout sequence s'

```
j =1;  s' [1]=s[1];
 for (i=1; i<|s|; i++){
 If (s'[i]≤s[i+1]) { s'[i+1]=s[i+1];   j=1;}
 else{
 mean =(s'[i]*j+s[i+1])/(j+1);
 for   (t=i-j+1;  t≤i+1;t++) s'[t]=mean;
 j++;
 } }
```

Isotonic regression of the histogram shown as Fig. 1, with the numerical values change shown as Fig. 5, and the red marks above the sequence as the abscissa of their corresponding components in the histogram. For convenience, data should be expressed with two decimal points.

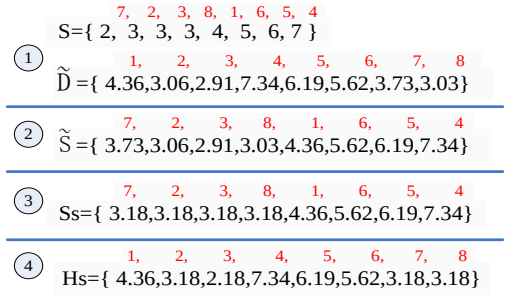


Fig. 5. Example of isotonic regression.

The isotonic regression algorithm of the corresponding sequences D and $\tilde{D}$ of the histogram in Fig. 1 follows several steps: 1. obtaining the ascending sort of histogram sequence without noise S and the histogram sequence $\tilde{D}$ with noise; 2. reconstructing $\tilde{D}$ to obtain $\tilde{S}$ according to the sort order (abscissa of the histogram) of sequence s; 3. isotonic regression of sequence S through, such as, Algorithm 1 to get the

isotonic sequence Ss; 4. reconstructing sequence *Ss* according to the abscissa of sequence $\tilde{D}$ to get sequence Hs, namely the final histogram sequence for release. Based on the order constraint of sequence *S*, the whole process introduces the minimum correction component to make the sort order of components of sequence $\tilde{S}$ unchanged before and after noise adding.

Based on the error calculation method above, the noise error through isotonic algorithm is Nerror(*D*, $\tilde{D}$)=3.6712, while the noise error without isotonic algorithm is Nerror(*D*, *Hs*)=2.1553, which obviously shows that isotonic algorithm reduces the final noise error of the histogram. Though isotonic regression incurs some error Error($\tilde{S}$, *Ss*)=0.4123, the price is relatively small compared to the noise error reduced by isotonic regression algorithm; moreover, the statistical sequence can be lengthy in practical application, so the error incurred by isotonic regression is negligible compared to the whole sequence data.

5. Experiment

The experiment is the simulation experiment of the privacy data release by histogram based on differential privacy-preserving model, and tests and verifies the effectiveness and reliability of isotonic regression of the histogram based on dynamic programming with specific index and data.

Experiment environment: CentOS6.5 as the operating system; Hive for data storage; Python2.6.6 as the programming environment; Python as the programming language.

Data source: Transactions of a company, from the website The Home of Data Science, including 349655790 records with each containing the commodity category. Because of the large data quantity, the numerical analysis is used in the experiment to detect and remove the singular dots, which ensures the reliability of the experiment result.

The differential privacy budgeε in the experiment is 1, 0.1, 0.01and 0.001. From the differential privacy definition, the more privacy budget, the higher requirement of privacy protection and thus the higher data processing cost required. The experiment mainly uses isotonic algorithm for post processing data based on dynamic programming noise to make the sequence unchanged before and after noise adding, which in some way ensures data accuracy and reduces release error of differential privacy data.

Whether the isotonic algorithm based on dynamic programming noise can reduce differential privacy data release is measured by data release, and the general release error is expressed as Terror. Because of the higher consumption of time and space by dynamic programming algorithm, five typical planning methods are used for experiment, and the groups number used are 1, 6, 10, 20 and 100 (which is 1, $n/15$, $n/10$, $n/5$, n, see reference [10]), and the experiment result is shown in Fig 6. E(H^*) is the error of the reconstructed histogram, namely the mentioned structure error $E(H^*)$, $E(Hm)$ is the error of the reconstructed histogram with noise Serror+Nerror, and $E(Hs)$ is the general error after isotonic regression.

It can be seen from each figure that regardless of the amount of privacy budget, as the groups planned increases, reconstruction error decreases, and in each reconstruction situation, data release error with isotonic regression is obviously less than the general release error without isotonic regression. Isotonic regression introduces a minimum correction value to a sequence before noise adding to ensure that the order of histogram sequence released is identical with that of histogram sequence before noise adding.

There is a significant trend of decline in error, on the one hand, because of isotonic regression algorithm, and on the other hand, because the minimum transformation in the transformation process is used to obtain the final isotonic result, and hence the reduction of release error of the histogram. Though correction error is incurred into isotonic transformation, the isotonic result improves the release accuracy of the whole histogram. Compared to the improvement degree of the histogram accuracy, isotonic correction error is negligible.

From the four figures, it can be seen that as privacy budget changes from *q* to 0.001, the reduction degree

of histogram error increases,. When privacy budget satisfies ε=0.001, the error range reduced by isotonic regression is maximum, hence the more significant the experiment result.

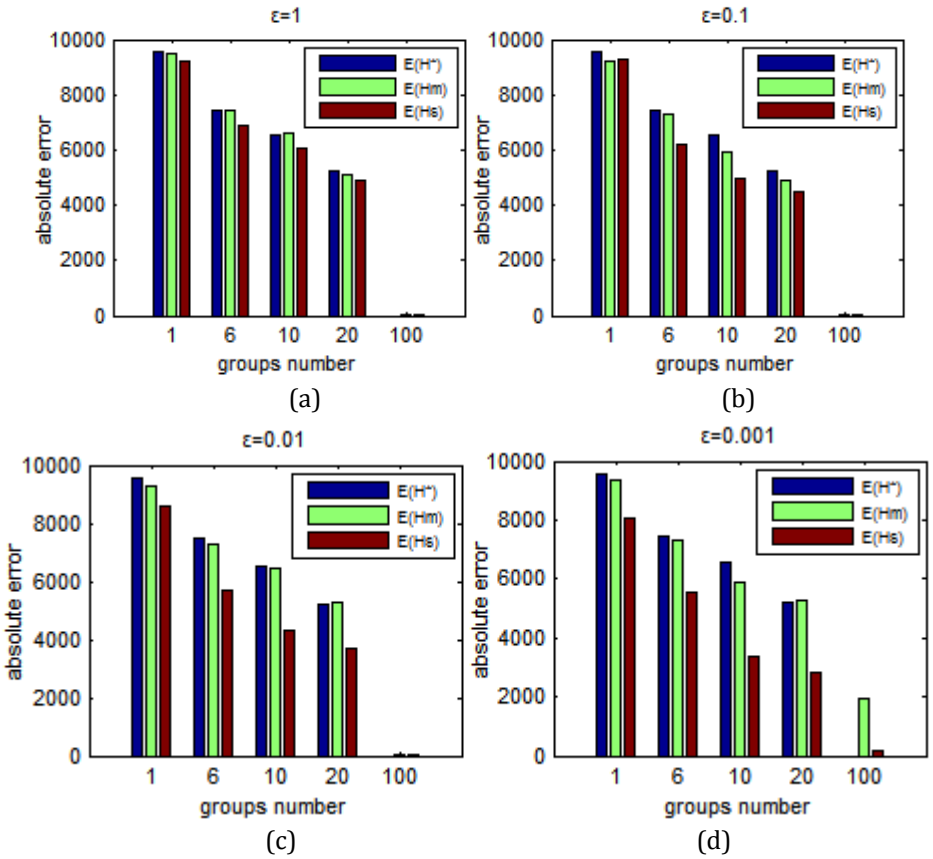


Fig. 6. Data release error.

6. Conclusion

Considering factors including privacy, safety and competitiveness, information is not suitable for direct publishing when data release, so data distortion through differential privacy model is desirable. This paper optimalizes the structure of data release via histogram with noise based on differential privacy model, and meanwhile, realizes isotonic regression on the basis of privacy data protection.

Prior research on this topic is focused on the reconstruction of the histogram, in order to reduce the range of privacy budget, instead of taking into account the histogram reconstruction error. Therefore, the release accuracy of the histogram is improved under the required differential privacy budget, and the sort order of the histogram sequence remains the same before and after data distortion. The accuracy of histogram release for privacy protection for the transaction data of a company is analyzed in later experiments, and the research project achieved the intended effect in view of the experiment result.

Moreover, with the increase of data quantity and more complicacy of data relationship, research on privacy data release poses more challenges. Besides data release by histogram, there are other methods, such as by space and by grid and so on. That will be the focus in the later research.

Acknowledgment

At the important moment of this paper is about to finalize, I should like to express my sincere gratitude to those persons who gave me kindness advice and support. I am greatly indebted to my advisor XiuJin Shi who guided the direction of my research and gave me valuable instructions. His effective advice, shrewd

comments have kept my passion on the thesis all way.

I would like to thank my team member for their encouragement and constructive suggestions. Their constantly help and Accompany make me full of power when I felt frustrated with this dissertation.

References

[1] Xiong, P., Zhu, T. Q., & Wang, X. F. (1989). A survey on differential privacy and applications. C*hinese Journal of Computers*, *37(1)*, 101-122.

[2] Sweeny. L. (2002). K-anonimity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, *10(5)*, 557-570.

[3] Wong, C. W., Li, J., Fu, W. C., *et al*. (2006). (Alpha, K)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. *KDD*, 754-759.

[4] Dwork, C. (2006). Differential privacy. *Proceedings of the 33rd International Colloquium on Automata*, *Languages and Programming* (pp. 1-12).

[5] Dwork, C. (2008). Differential privacy: A survey of results. *Lecture Notes in Computer Science*.

[6] Zhang, X. J., & Meng, X. F. (2014). Different privacy in data publication and analysis. *Chinese Journal of Computers*, *4*, 927-949.

[7] Yaroslavtsev, G., Cormode, G., Procopiuc, C. M., *et al*. (2013). Accurate and efficient private release of datacubes and contingency tables. *Proceedings of the 29th International Conference on Data Engineering* (pp. 745-756).

[8] Dwork, C., Mcsherry, F., Nissim, K., *et al*. (2006). Calibrating noise to sensitivity in private data analysis. *Proceedings of the Theory of Cryptography Conference* (pp. 265--284).

[9] Jagadish, H. V., Poosala, V., Koudas, N., *et al*. (1988). Optimal histograms with quality guarantees. *Vldb*, 275–286.

[10] Xu, J., Zhang, Z., Xiao. X., *et al*. (2013). Differentially private histogram publication. *Vldb Journal*, *22(6)*, 32-43.

[11] Hay, M., & Miklau, G. (2010). Boosting the accuracy of differentially-private histograms through consistency. *Proceedings of the International Conference on Very Large Data Bases*, *3(12),* 66-69.

[12] Barlow, R. E., & Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, *67(337)*, 140-147.

XiuJin Shi is a professor of computer science and technology at Donghua University. He main researches digital fashion, database applications, software engineering theory and technology. He is one member of Shanghai Computer Society and the Association of Shanghai Computer open systems. He has 6 years of experience in teaching and guiding projects for undergraduate and post graduate students. He has to his credit 15 publications in national/international conferences and journals.

Ling Zhou is a graduate student from Donghua University, major in software engineering. Her master's research subject is information security and privacy protection. During this period, she received basic knowledge through her own efforts, and contributed to the study.