Research of Differential Privacy Protection on Social Network

Jian Zhang, Changlun Zhang, Shihui Chen and Shupeng Wu

School of Science, Beijing University of Civil Engineering and Architecture, 100044, Exhibition Hall Road, Xicheng District, Beijing, China.
Email: zhangjian@bucea.edu.cn; zclun@bucea.edu.cn

Abstract. Rapid development has been witnessed in social network. As a consequent, much more people are involved in getting it, sharing their own information to enlarge their social circle. In this paper, the security problem, which exists in privacy preserve in social network and process users’ private information via Differential Privacy Histogram Release is focused on. The method can process user's privacy information on the premise of guaranteeing the availability of information, making the information released after publication different from the information before publication, thus making the user's privacy information not leak. Shown by theoretical analysis and research results, safe release of private information can be guaranteed if processed by this method.

1. Introduction
Social network seems getting much more familiar to us for the sake of its great development. While more and more user getting involved in it, they tend to put on their personal information. However, unnecessary damage may be brought to users by some hostile cyber hackers since the information probably relate to something private. As for the reason, it is obvious that these users’ information exactly reflects their bias or interest in some aspects. And through analysis, users’ private information like interest and disposition can be available to dig out their characteristic in terms of social acquaintance or group and so on, which results in the disclose of a part of users’ privacy. Thus, it’s a necessity to protect users’ published information and keep their safety. Differential Privacy Preserve is a kind of privacy preserving technique based on distortion process of data. It achieves data distortion mainly by adding noise to original data. And simultaneously it still keeps the attributes of data which maintain part of statistical characteristics of data, and make it much more available to further research like data mining. In this essay, we apply Differential Privacy Histogram Release in the data release process. By processing these unreleased data, we achieve the protection of users’ privacy.

2. Relative Work
The problem of statistic privacy protection is initially come up with by Statistician Dalenius in late 1970s. It considered that in order to protect private information in database, any arbitrary individual’s accurate information is not available when any users (including legal users and potential attacker) try to access database. Early technology for privacy protection is K-anonymity raised by Sweeney in 2002, in which generalization to attribute of data is committed to realize preservation of dataset privacy. In 2006, Machanavajjhala found a great vulnerability existing in K-anonymity and thus, came up with another way of privacy preservation based on 1-Diversity Principle. Through further generalization of already generalized attribute, model showed better performance on privacy preservation. During the same year, targeting at the problem of privacy disclose in statistical databases, Dework put forward a model of privacy preservation, Differential Privacy Preserve. Extracted from it, firstly the model strictly define acknowledge background held by attacker. Even under this situation where attackers
would be probably accessible to all information except the real target, Differential Privacy Preserve still keep the preservation of privacy. Secondly, it introduces strict mathematics theory to this field, producing a way of calibrating how Differential Privacy Preserve performed. Thus, this method appealed a great attention from the beginning. During the same year, (see Reference 7), Dwork introduced $\varepsilon$ - Differential Preserve and gives a mechanism providing Differential Privacy Preserve, Laplace Mechanism. By perturbing the true answer to a database query by the addition of a small amount of Laplace or exponentially distributed random noise, it achieved protection of query result. However, on some occasions because some query functions held a larger global sensitivity, added noise became even larger, which resulting in impaired usability of queried result. In 2007, Nissim defined partial sensitivity of query functions, which improved the usability of information simultaneously besides Differential Preserve. In the same year, McSherry proposed another mechanism providing Differential Privacy Preserve, Exponential Mechanism. It mainly suit for non-numeric dataset. In 2011, Dwork together with others came up with $(\varepsilon, \delta)$ - Differential Privacy Preserve. During process of perturbing random noise it used Gaussian Distribution instead. Since Differential Privacy Preserve offering a proved, quantitative protection, it has already extensively applied to many fields.

3. Mechanism of Differential Privacy Preserve

3.1. Fundamental Concept

Definition 1$^{[10]}$(Differential Privacy)

Supposing a random algorithm $M$, and $P_M$ is the set comprised of every possible output of $M$. For any two adjacent dataset $D$, $D'$ and any subset $S_M$ of $P_M$, if algorithm $M$ satisfy:

$$P_r[M(D) \in S_M] \leq \exp(\varepsilon) \times P_r[M(D') \in S_M]$$

Then algorithm is said to provide $\varepsilon$ - Differential Privacy Preserve, and parameter $\varepsilon$ is called privacy preserve budget.

From the definition of Differential Privacy, we can find that it theoretically ensures algorithm $M$ satisfy $\varepsilon$ - Differential Privacy Preserve. Yet, we need to introduce noise mechanism to achieve Differential Privacy Preserve on dataset.

3.2. Mechanism of Differential Privacy Preserve

Dataset includes data which can be numerical or non-numerical. So, we should take different mechanism of Differential Privacy Preserve for different types of data. Laplace Mechanism for numerical data, proposed by Dwork in 2006 and Exponential Mechanism for non-numerical data, proposed by McSherry in 2007 are depicted as followed.

(1) Laplace Mechanism

Definition 2$^{[7]}$(Laplace Mechanism)

For given dataset, if there exists function $f : D \rightarrow R^d$ of which global sensitivity is $GS_f$, if algorithm $M$ satisfy:

$$M(D) = f(D) + \text{Lap}(GS_f / \varepsilon)$$

Then it is said to providing $\varepsilon$ - Differential Privacy Preserve. $\text{Lap}(GS_f / \varepsilon)$ Is random noise obeying Laplace Distribution with scale parameter $GS_f / \varepsilon$.

(2) Exponential Mechanism

Definition 3$^{[11]}$(Exponential Mechanism)

Supposing a random algorithm $M$, its input dataset. $D$, algorithm output a realistic data $r \in \text{Range}$. $q(D, r)$ Is the usability function of $r$ to evaluate the performance of output $r$ and $\Delta q$ is the sensitivity of function $q(D, r)$.
If algorithm $M$ choose to output $r$ from Range with the possibility which is proportional to $\exp\left(\frac{\varepsilon q(D, r)}{2\Delta q}\right)$, then the algorithm is said to provide $\varepsilon$-Differential Privacy Preserve.

### 3.3 Differential Privacy Histogram Release

Histogram can intuitively display the distribution of some data and is sensitive to the change of data. The increase or decrease of data will directly affect the change of histogram shape. In addition, the histogram is more convenient to count. Under the premise of differential privacy, it can easily calculate the counting query and the sensitivity of the query. Therefore, it's capable to adopt Differential Privacy Histogram Release for releasing data. For the histogram release of the data, the data is firstly divided into $m$ isometric interval segments according to the characteristics of the dataset, and count the number of data in each interval segment. In order to protect the data, Laplace noise needs to be added separately to the data frequency of each section so that there is a certain difference between the frequency before and after the data is released. Finally, we can calculate the frequency difference before and after the release of data and privacy protection level.

![Figure 1. Histogram to add noise before and after](image1)

$$H_i = H_i + \text{Lap}(\Delta f / \varepsilon) \quad i = 1, 2, \ldots, 7$$

### 4. Privacy Protection Application Based on Differential Privacy in Social Network

In social network, users should fill their personal information after registration, such as personal statistics, personal labels of Microblog users shown in Figure 2, the information is all associated with user's private information, and has a certain privacy disclosure risk before directly public release if without processing. Therefore, privacy protection is necessary to ensure that any privacy information is not disclosed before releasing the information.

In general, age and gender information are an important attribute of the user. Therefore, it's necessary to process data while querying the two properties above. And we use the method of Differential Privacy Histogram Release to process the user's gender and age comprehensively so that the queried data cannot be analysed to get the user's specific age and gender.

![Figure 2. User Registration of Personal Information](image2)

| Mobile number | 100218 | 1990 1 |
|---------------|--------|--------|
| Nickname      | 100232 | 1994 1 |
| Profile       | 100248 | 1983 1 |
| Shipping address | 100267 | 1985 1 |
| Education     | 100283 | 1978 2 |
| Domain Name   | 100303 | 1987 1 |
|               | 100338 | 1983 1 |

![Figure 3. The Data Set Form](image3)
4.1 Preparation Stage

Presuming a dataset $H$ which contains users’ ID, Age for ages, Sex for gender, its formation is the same as Figure 3. The dataset is comprised of users of different generation and different genders apparently. First, we divide these data into $w$ pieces of isometric original histogram with $w$ data records according to generation and genders. $H = \{H_1, H_2, ..., H_w\}$, among it each $H_i$ contains one user’s age and gender. That is $H_i = \{\text{Age}_i, \text{Sex}_i\}$. Through analysis of this dataset, we use number 1 and 2 represents users’ gender, and we found in this specific dataset, all users’ ages locate between $[15, 60]$. So, we divide ages set into nine ranges $\{[15, 20], [21, 25], [26, 30], [31, 35], [36, 40], [41, 45], [46, 50], [51, 55], [56, 60]\}$. Now we need 18 isometric histograms to count users’ number. This unprocessed histogram release would straight leak users’ privacy, thus we need to process information in the histogram, which is described as followed ‘Privacy Protection Stage’.

4.2 Privacy Protection Stage

Considering that Differential Privacy Histogram Release relates to one-dimensional data, we need to transform users’ ages and genders in $H_i$ to a data form with only one attribute. What’s more, transformed data must be available for further search and reversible to original data by some way. So that this data possesses equal security as well as high usability. The process of data is described as Figure 4 and Figure 5. Then we take some measures to mix up age information together with gender information and still use the former age ranges. Now we get the one-dimensional data including age and gender information, which would be able to use in histogram release.

![Figure 4. The original data partition](image)

![Figure 5. Processed data partition](image)

Through analysis above, while using Differential Privacy Histogram Release analyzing users’ ages and genders, firstly we should divide data into $n$ isometric ranges, forming $n$ parts of isometric original histograms $H = \{H_1, H_2, ..., H_n\}$, then adding Laplace noise to counts of each range of the histograms in order to interfere in real data. So the present counting histograms become $H = \{H_1, H_2, ..., H_n\}$, among which $H_i = H_i + \text{Lap}(1/\epsilon)$. Laplace Distribution of the parameter $1/\epsilon$. $\epsilon$ Represents privacy preserve budget. However, this method failed in queried results for too larger accumulated noise, which means low usability of the answer $H$. Consequently, Histogram Release above need improving when releasing a large amount of data. Practically, we combined some adjacent ranges when there occurred some usability in queried results. As is shown in Figure 6, this improvement heightened the length of range which is available to query and improve usability of the queried results.

For example, viewing from Figure 6 and Figure 7 which are the simulation of a partial data of social network, Figure 6 shows counts of users’ ages and genders using original histogram released data, under the condition of noise $\text{Lap}(\Delta f / \epsilon)$ added into each range of histogram, accounting...
9* Lap(Δf / ε) for all noise. Figure 7 depicts results of partition histogram release. Frequency of each partition range is the average of frequency of its original range under the condition of the same noise Lap(Δf / ε) added into each range of histogram, accounting 3* Lap(Δf / ε) for all noise. There inevitably exists error between frequency of original histogram and that of partition histogram, but it contributes less effect to queried results.

![Figure 6. The original histogram](image1)

![Figure 7. Partition histogram](image2)

4.3 Release Stage
In the premise of adequate usability of released data, as to data needing release, we deal with them with differential privacy process, and employ Differential Privacy Histogram Release while need to release data. When processing dataset of small amount, we could directly use original Differential Privacy Histogram Release to release data. But as for larger one, we should take partition process onto the dataset, and then apply Differential Privacy Histogram Release to ensure the security as well as availability for further research.

5. Privacy Preserve Effect Analysis
While we using the Differential Privacy Histogram Release to implement privacy preserve on data, privacy preserve budget ε is such a critical parameter. Value of ε should be determined via analysing users’ specific requirement so as to balance security and usability of released data. Generally, ε should take a smaller value since larger one may result in impairing privacy preserve. And sensitivity of query function Δf affect privacy preserve level to some extension as well.

Using statistics from Tecent Microblog (referred in KDD competition, 2012), we applied Differential Privacy Histogram Release on it for privacy preserve analysis. Considering the simplicity, we only extracted age and gender information of the entire dataset. As two-dimensional data was given, we took a first step to process data. We combined the gender information into age information and then convert the age range [15, 60] to combined data range [1, 120]. Next, we carried out privacy preserve analysis with combined data, mainly in terms of querying users’ numbers in each range, that is \( f_i = \text{count}(i) \).

While carrying out Differential Privacy Histogram Release for data analysis, we found remarkable impact originated from privacy preserve budget ε and sensitivity of query function Δf. And in the following part, via simulation, we would discuss specific impact on privacy preserve brought by these two parameters with Microblog users’ age and gender information.

5.1 Effect Brought by Privacy Preserve Budget
When referring to effect brought by privacy preserve budget, firstly we may assume that Δf = 1 and ε value from \( \{0.01, 0.1, 1\} \) to evaluate the effect. Given that data from dataset range in \([0, 120]\), we mainly analyse histograms in five ranges, whose counting numbers respectively are \( \{1, 8, 12, 24, 120\} \).

Figure 8 (a), (b), (c) listed as followed indicate, under the different value of ε, which are 0.01 0.1 and 1 respectively, research results in different range k by Different Privacy Histogram Release. And k
values in \{1, 8, 12, 24, 120\}. Viewing from the figures, we may easily find that under the same privacy preserve budget, error differs from each range, but still shares the same magnitude. For instance, as is depicted in Figure 8(a), when privacy preserve budget take the value of 0.01, each separate range shares the same magnitude of mean square error around e+05. Moreover, under different privacy preserve budgets, though differing from the same release range, magnitude of error varies with exponential (base 10) growth. For instance, privacy preserve budget valuing as 0.01 0.1 and 1, take the according magnitude of e+05 e+04 and e+02. From comprehensive analysis with these three figures, when we assume \( \Delta f = 1 \), with the increase of privacy preserve budget \( \varepsilon \), added Laplace noise decrease and lower the error between query result and real result. This is mainly because with the increase of \( \varepsilon \), Laplace noise added into the original query dataset \( \text{Lap}(1/\varepsilon) \) get decreasing as well, which, thus, lower the privacy preserve level and gradually increase usability reversely. Concluded from these, we may reach a conclusion that as privacy preserve budget \( \varepsilon \) getting smaller, privacy preserve presents more remarkable effect via Differential Privacy Histogram Release.

![Image](image1.png)

(a) \( \varepsilon = 0.01 \)  
(b) \( \varepsilon = 0.1 \)  
(c) \( \varepsilon = 1 \)

Figure 8. The influence of privacy preserve budget \( \varepsilon \) on privacy preserve

5.2 Effect Brought by Sensitivity of Query Function

When referring to effect brought by privacy preserve budget, we assume \( \varepsilon = 0.01 \), sensitivity of query function \( \Delta f \) values in \{0.1, 1, 10, 50\} to discuss its impact on privacy preserve. Given that data from dataset range in \[0, 120\], we also mainly analyse histograms in five ranges, whose counting numbers respectively are \{1, 8, 12, 24, 120\}.

Figure 9 (a), (b), (c), (d) listed as followed indicate, under the different value of \( \Delta f \), which are 0.1 1 10 and 50 respectively, research results in different range k by Different Privacy Histogram Release. And k values in \{1, 8, 12, 24, 120\}. Viewing from the figures, we may easily find that under the same query function sensitivity, error differs from each range, but still shares the same magnitude. For instance, as is depicted in Figure 9(a), when query function sensitivity takes the value of 0.1, each separate range shares the same magnitude of mean square error around e+02. Moreover, under different query function sensitivity, though differing from the same release range, magnitude of error varies with exponential (base 10) growth. For instance, query function sensitivity valuing as 0.1 1 10 and 50, take the according magnitude of e+02 e+04 e+06 and e+07. As is shown in Figure 9, when we assume \( \varepsilon = 0.01 \), with the increase of query function sensitivity \( \Delta f \), added Laplace noise gets larger and increase the error between query result and real result. This is mainly because with the increase of \( \Delta f \), Laplace noise added into the original query dataset \( \text{Lap}(\Delta f / 0.01) \) get larger as well, which, thus, promote the privacy preserve and gradually decrease usability reversely. Concluded from these, we may reach a conclusion that as query function sensitivity \( \Delta f \) getting larger, privacy preserve presents more reliable privacy preserve via Differential Privacy Histogram Release.
6. Summary
In this essay, we illustrate process of queried data for users’ privacy preserve based on Differential Privacy Histogram Release. We add noise which obey Laplace Distribution \( \text{Lap}(\Delta f / \varepsilon) \) in frequency of data histogram to disturb original data on network and obtain error between data released before and after. And we also analyse quantitative error between queried result and original data under condition of different privacy preserve budget \( \varepsilon \) and different sensitivity of query function \( \Delta f \).

Indicated by theoretical analysis and research results, Differential Privacy Histogram Release is efficient in protecting data on network, and realize secure release of data on network.

7. Acknowledgement
This work was supported by scientific research fund project in Beijing University of Civil Engineering And Architecture (KYJJ2017035).

8. References
[1] Dalenius T. Towards a methodology for statistical disclosure contro1. Statistik Tidsskrift, 1977, 15: 222-449.
[2] Xiong Ping, Zhu Tian-Qing, Wang Xiao-Feng. A Survey on Differential Privacy and Applications [J]. Chinese Journal of Computers, 2014, 37(1): 101-122.
[3] Sweeney L. K-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10 (5): 557-570.
[4] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10 (5): 571-588.
[5] Machanavajjhala A, Gehrke J, Kifer D, et al. 1-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006. 24-35.
[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Theory of Cryptography, ser. Lecture Notes in Computer Science, S. Halevi and T. Rabin, Eds. Springer Berlin / Heidelberg, 2006, 3876: 265-284.

[7] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, “Our data, ourselves: Privacy via distributed noise generation,” in Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, ser. EUROCRYPT ’06. Berlin, Heidelberg: Springer-Verlag, 2006: 486-503.

[8] Dwork C. The promise of differential privacy: A tutorial on algorithmic techniques. In: Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. Palm Springs, USA, 2011: 1-2.

[9] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency//Proceedings of the 36th Conference of Very Large Databases (VIDB). Istanbul, Turkey, 2010: 102/1032.

[10] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011, 54(1): 86-95.

[11] F. McSherry and K. Talwar, “Mechanism design via differential privacy,” in Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, ser. FOCS’07. Washington, DC, USA: IEEE Computer Society, 2007: 94-103.