ATLANTIS PRESS
Advances in Computer Science Research, volume 50
4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016)

The Study of Data Publishing Technology based on the Differential Privacy in Social Networks

Nan Ning[1, a,*], Changlun Zhang[1, b], Zhanyong Jin[2,c], Zhan Yu[1, d]

[1]School of Science,Beijing University of Civil Engineering and Architecture, Beijing,100044, Chain

[2]School of Economics and Management of Engineering, Beijing University of Civil Engineering and Architecture, Beijing,100044, Chain

[a]email: ningnan@stu.bucea.edu.cn, [b]email: zclun@bucea.edu.cn,

[c]email: jinzhanyong@ bucea.edu.cn, [d]email: 2107010415004@stu.bucea.edu.cn.

Keywords: Privacy Preservation; Differential Privacy; Social Networks

Abstract. With the increasing prevalence of social network, research on privacy preserving data publishing in the social network has received substantial attention recently, and the recent emergence of differential privacy has shown great promise for rigorous prevention of information publishing. In this paper, we applied the differential privacy to protect the user information during the data publishing and provided a holistic solution for data publication. In addition, we also explored the influence caused by the query function sensitivity and the privacy preserving budget. The results show that the privacy protection degree increases with the increasing of the privacy preserving budget, while decreases with the increasing of the query function sensitivity.

1. Introduction

With the development of social networks, a large volume of user data has been generated, which enables a wide spectrum of data analysis tasks. Consequently, the data privacy preserving plays an important role in data publishing.

The major challenges of applying privacy data protection technology include data privacy preserving degree and data utility, and various researches have been developed to respond to these challenges. A pioneering step in this direction was taken by Latanya, who considered k-anonymity technology to protect the privacy data, which makes the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release [1]. Since the k-anonymity technology is subject to the uniform attack, scholars proposed the l-diversity technology, dividing each sensitive attribute generated by the k-anonymity technology into at least l different values, which can resist the uniform attack [2]. At the same time, considering privacy protection only against a specific attack model makes it difficult to obtain the realistic need. As a result, Dwork et al. proposed a differential privacy protection model by inserting the noise that is related to the query function into the original data in order to make the privacy information can't be stolen [3]. Since the solid mathematical foundation and strong attack model of the differential privacy, it has become a hot topic of privacy research, and has been widely used in many fields, such as machine learning, secure communication and social networks.

The application of differential privacy in social networks is divided into two parts, some researches focus on the node's differential privacy protection and others focus on the edge differential privacy protection. The node's differential privacy technology is to delete a node and the links connected to this node in the graph, so that all individual privacy information will not be stolen, although the query function is limited. The edge differential privacy technology is to delete any k edges in the graph, through which the data availability increases and the protection degree decreases, however, it is enough to use in some field so that the edge differential privacy be used more extensively. Task et al. presented the concept of out-degree differential privacy protection, which means deleting a node and the node out-degree. As same as the edge differential privacy technology, this method increases the data availability [4]. Furthermore, the differential privacy can

also be used for social network analysis, which mainly has the following several methods: the degree distribution query, triangle query, k-triangle count query, k-triangle count query, calculating clusters coefficient and calculating the weight of edges [5-8]. These above methods can be used to protect various data information aspects, such as scalability, usefulness and utility. However, these research most be used to prevent the social network structure information from stealing.

In this paper, we introduce a data publishing method based on the differential privacy in the social networks, in which inserting noise satisfied Laplace distribution to the publishing data causes the data cannot be stolen. Furthermore, we explore two parameters influence on the proposed method, and the theoretical results show that during the data publishing process, the differential privacy technology is efficient to the privacy protection.

2. Preliminaries

In this section, we will introduce the preliminaries used in this paper.

2.1 Differential Privacy

Def 1: A privacy mechanism A gives ε-differential privacy, if for any database D_1 and D_2 s. t. $|D_1 \Delta D_2| = 1$, and for any possible output $O \in Range(A)$, $\Pr[A(D_1) = O] \le e^{\varepsilon} \times \Pr[A(D_2) = O]$, where the probability is taken over the randomness of A [9].

2.2 Noise mechanism

Noise mechanism is one of the main technologies to achieve the differential privacy protection, which is divided into two kinds: the Laplace noise mechanism and the exponential mechanism. Laplace noise mechanism is mainly applied to the data set which is the numerical model, while the exponential mechanism is applicable to the situation when the data set data is the abstract entity data. Due to the data used in this paper are numerical, it is mainly used Laplace noise mechanism.

Def 2: For any function $f : D \to R^d$, the global sensitivity of f is $\Delta(f) = \max \| f(D_1) - f(D_2) \|_1$ for all D_1, D_2 s. t. $|D_1 \Delta D_2| = 1$ [10].

Th 1: For any function $f : D \to R^d$ over an arbitrary domain D, the mechanism $A(D) = f(D) + Lap_1(\Delta f / \varepsilon), \cdots, Lap_d(\Delta f / \varepsilon)$ gives ε-differential privacy.

Here, $Lap_i(\Delta f / \varepsilon)$ $(1 \le i \le d)$ is independent Laplace variable [11].

2.3 Combination characteristic of differential privacy

Th 2: Let A_i each provide ε_i-differential privacy.A sequence of $A_i(D)$ over the database D provides $\sum \varepsilon_i$-differential privacy [11].

Th 3: Let A_i each provide ε_i-differential privacy.A sequence of $A_i(D_i)$ over a set of disjoint databases D_i provides $\max(\varepsilon_i)$-differential privacy [11].

3. Method

This section will describe the method in detail, as well as our assumptions and conclusions.

We assume that the existing data set is T, and each record contains n different attributes $X_1, \cdots, X_n$, $|X_i| (1 \le i \le n)$ represents the number of each attribute. As the database is large in reality, we need divide the database into disjoint sub database firstly. Then, query on the sub database.

Algorithm1: Data Publishing

Input: Data set T.

Output: The query results $F(T)$ after inserting noise.

1) Initialize the value of privacy preserving budget.
2) Group the new data set T.
3) Replace the data value in the sub-database T_i to get a new database.

4) Sum the data in each field to obtain the results of the query $f(x)$.
5) Calculate the global sensitivity Δf of the query function f.
6) Generate Laplace sequence which obeys location parameter is 0 and the scale parameter is $\Delta f / \varepsilon$.
7) Select $Lap(\Delta f / \varepsilon)$ from the Laplace sequence randomly.
8) Calculate the query result after inserting the noise $F(x) = f(x) + Lap(\Delta f / \varepsilon)$.

Conclusion 1: The sensitivity of the query function in the above algorithm is 1.

Proof: According to the above algorithm, in the implementation process of the algorithm, $f(T_i)$ and $f(T_i^{'})$ respectively represent the query results in the data subsets T_i and $T_i^{'}$, which only different in one data. And the global sensitivity for the query function is one of the following four states:

State One:

$$\begin{cases} f(T_i) = 0 \\ f(T_i^{'}) = 0 \\ \left\| f(T_i) - f(T_i^{'}) \right\| = 0 \end{cases} \quad (1)$$

State Two:

$$\begin{cases} f(T_i) = 0 \\ f(T_i^{'}) = 1 \\ \left\| f(T_i) - f(T_i^{'}) \right\| = 1 \end{cases} \quad (2)$$

State Three:

$$\begin{cases} f(T_i) = 1 \\ f(T_i^{'}) = 1 \\ \left\| f(T_i) - f(T_i^{'}) \right\| = 0 \end{cases} \quad (3)$$

State Four:

$$\begin{cases} f(T_i) = 1 \\ f(T_i^{'}) = 0 \\ \left\| f(T_i) - f(T_i^{'}) \right\| = 1 \end{cases} \quad (4)$$

From the above four kinds of state, we can get $\Delta f \le \max \left\| f(T_i) - f(T^{'}_i) \right\| = 1$.

Conclusion 2: The algorithm satisfies the ε – differential privacy.

Proof: Due to the fact that the data set T is divided into k disjoint sub databases, by theorem 2, the algorithm which meets with ε – differential privacy is still meeting with ε – differential privacy in disjoint data sets.

4. Discussion

In this part, according to the actual situation, we assume the data consists of eight field properties, which are user name, age, gender, date of birth, occupation, qualifications, hobbies, micro-blog numbers, and each field properties have 50 data. We would like to protect the privacy of the user age and our query function used in the model is count query. After dealing with the database, we can get the following two tables:

Table 1 Processed database table

Age state	After data processing
0-10	1
10-20	2
20-30	3
30-40	4
More than 40	5

Table 2 Number of each age state table

Age	1	2	3	4	5
Number	1	12	16	10	11

According to the theoretical analysis of evolution model, the parameters of Δf and ε has a significant effect on the results of privacy protection. In this section, we will discuss if privacy preserving budget and query function sensitivity impact on the results of privacy protection.

4.1 Influence of privacy protection budget ε

When we discuss the effect of privacy protection budget ε on the protection of privacy, we assume $\Delta f = 1$, and ε respectively take (a) $\varepsilon = 0.01$, (b) $\varepsilon = 0.1$, (c) $\varepsilon = 0.5$, (d) $\varepsilon = 1$.

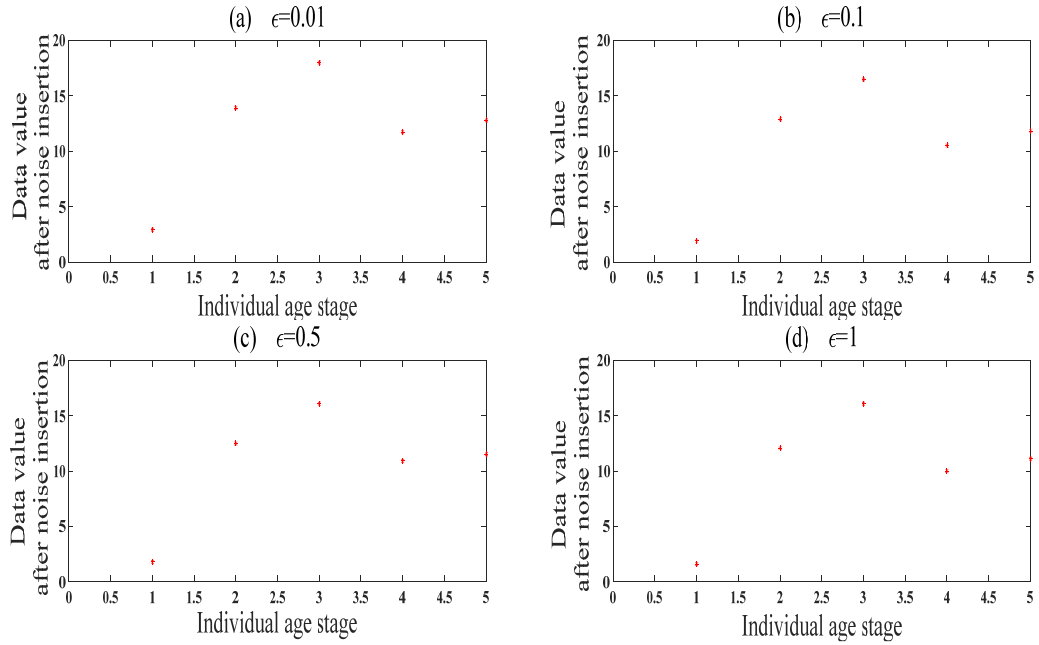

Figure 1 Influence of ε

As shown in Figure 1, the value of ε is greater than 0, and the value of Δf is 1. With the increasing value of ε, the output of the algorithm decrease obviously, and the difference between the original query results and the experimental results also decreases. The reason is that, with the increase of ε, the noise which should be inserted into the query results is reducing. In this case, however, the protection of the original query results also reduced. The availability of the data has been affected. This shows that the smaller the value of privacy budget ε is, the higher the degree of privacy protection provided by this algorithm is.

4.2 Influence of the sensitivity of the query function Δf

When we discuss the effect of the sensitivity of the query function Δf on the protection of privacy, we assume $\varepsilon = 0.01$, and Δf respectively take (a) $\Delta f = 0.1$, (b) $\Delta f = 1$, (c) $\Delta f = 10$, (d) $\Delta f = 50$.

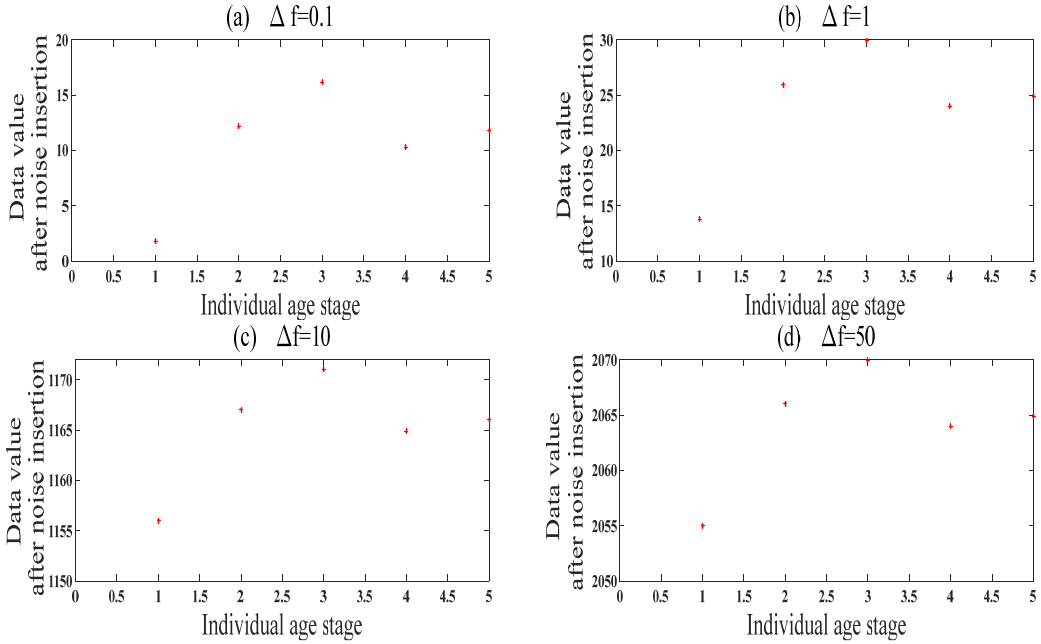

Figure 2 Influence of Δf

As shown in Figure 2, the value of ε is 0.01. With the increasing value of Δf , the output of the algorithm continues to increase ,and the gap between the result which the noise has been added and the original result is in the continuous increase. This phenomenon shows that, with the increasing value of Δf , the noise added to increase, and the degree of privacy protection is improving. But this algorithm has caused the problem of poor availability of the original data.

5. Conclusion

This paper gives a data publishing method based on the differential privacy in social networks. The method makes the user's information unable to leak by adding noise which obeys Laplace distribution. In addition, we simulate the influence of two characters which are the privacy budget and the query function sensitivity on the publishing method. Theoretical and simulation results show that the method success in releasing the user data safely. And the privacy protection degree increases with the increasing of the privacy preserving budget, while decreases with the increasing of the query function sensitivity.

Acknowledgement

In this paper, the research was sponsored by the National Nature Science Foundation of Chain (Project No. 61401015 and No. 7154100034) and Beijing Municipal Education Commission on Projects (SQKM201510016013).

References

[1] Latanya Sweeney. K-anonymity: A model for protecting privacy. International Journal of Uncertainty[J]. Fuzziness and Knowledge-Based Systems, 2002(10) 557-570.

[2] Ashwin Machanavajjhala, Daniel Kifer, Johannes Geherke. L-diversity: Privacy beyond k-anonymity[J]. ACM Transaction on Knowledge Discovery from Data, 2007.

[3] Dework Cynthia, Mcsherry Frank. Calibrating noise to sensitivity in private data analysis[J]. Springer Berlin, 2006.265-284.

[4] Friedman Arik, Schuster Assaf. Data mining with differential privacy[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010.493-502

[5] Mohammed Noman, Chen Rui, Fung Benjamin C.M. Differentially private data release for data mining[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.493-501.

[6] Li Chao, Hay Michael, Rastogi Vibhor. Optimizing linear counting queries under differential privacy[C]. Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2010.123-134.

[7] Task Christine, Clifton Chris. A guide to differential privacy theory in social network analysis[C]. Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, 2012.411-417.

[8] Li Chao, Hay Michael, Miklau Gerome. Accurate estimation of the degree distribution of private networks[C]. Proceedings of the 9th IEEE International Conference on Data Mining, 2009.169-178.

[9] Dework Cynthia. Differential privacy[C]. Processing of the 33rd International Colloquium on Automata, Languages and Programming, 2006.1-12.

[10] Nissim Kobbi, Raskhodnikova Sofya, Smith Adam. Smooth sensitivity and sampling in private data analysis[C]. In Proceedings of the 39th Annual ACM Symposium on Theory of Computing, 2007.75-84.

[11] Mcsherry Frank. Privacy integrated queries: An extensible platform for privacy preserving data analysis[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, 2009.19-30.