Privacy Preservation in Skewed Data using Frequency Distribution and Weightage (FDW)

[†]Sabah Suhail and [*]Choong Seon Hong

Department of Computer Engineering, Kyung Hee University, Yongin, 446-701 Korea

{[†]sabah, [*]cshong}@khu.ac.kr

Abstract

Many organizations collect and distribute personal data for a variety of different purposes, including demographic and public health research. Disseminating such critical personal data for analysis purposes requires data publishers to enforce privacy-preserving models. The syntactic privacy models for static data publishing have only discussed datasets that have distinct values of sensitive data leaving the skewed datasets unaddressed. In this paper, we focus on skewed datasets, and follow the anatomy rule of releasing all quasi-identifiers and sensitive values directly in two separate tables. However, to preserve the privacy of individuals in skewed data we use frequency distribution scheme instead of count in sensitive table. Moreover, weighted column is used to enable data analyst from correctly understanding the data distribution inside each QI-group.

1. Introduction:

Privacy models and algorithms in Privacy Preserving Data Publishing (PPDP) facilitate the data publishers in achieving a trade-off between privacy and utility via sanitization mechanisms. The pressing question is how to design a perfect privacy preserving sanitization mechanism which outputs extremely useful data? To answer this question, Sweeney [1] proposed a well-known syntactic privacy definition, k-anonymity, for preventing linking attacks using quasi-identifiers. While k-anonymity privacy model protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. Therefore, it does not take into account the attributes that contain highly sensitive information (e.g., patient diagnosis, salary, occupation etc.). Machaanavajjhala et al. [2] highlight this problem of k-anonymity model and propose a privacy model, l-diversity, that focus on protecting the association between individuals and these sensitive values. l-diversity model, ensures that there are at least one distinct sensitive values for the sensitive attribute in each equivalence class. Both of these techniques preserve privacy by generalization QI values. However, a serious drawback of generalization is that, when the number dimensions of QI attributes is large, any generalization necessarily loses considerable information in the microdata [3], due to the "curse of dimensionality". Specifically, in high dimensional spaces, each generalized value is always an exceedingly wide interval, resulting in useless published table for research purpose.

To resolve this issue, Xiao et al. [4] introduce a technique which preserves both privacy and correlation in the microdata, and hence, overcomes the drawbacks of generalization. However, if we apply the anatomy technique on dataset containing skewed data, then it may be going to reveal the exact count of patients suffering from a particular disease in an equivalence class.

1.1 Rationale of frequency distribution in skewed data

Realistically there are medical datasets which are containing information about different forms of any particular disease or any particular semantic-specific information. For instance, dataset of chronic diseases (cancer, HIV) or dataset holding information about blood group of individuals. The semantic of sensitive attributes in these datasets are somehow identical apropos to the disease. To preserve the privacy of individuals in skewed datasets, generalization or anatomy may not be able to give optimal results based on utility-privacy tradeoff.

Consider a scenario in which an adversary, Mallory has some auxiliary information about Bob including his personal details (age 23 and zipcode 11k). Also she only knows that Bob is suffering from disease *d*. If we apply anatomy in this scenario, the sensitive anatomized (Table 2) is clearly revealing that in equivalence class 1, the patient is having angina through count column.

Data skewness can help adversary to infer with high assurance that victim has disease *d* by looking at the highest count in any particular equivalence class *E*. Hence it's difficult to prevent the attribute/identity disclosure of any individual and as a result the privacy of released dataset is highly questionable.

2. Proposed Methodology:

For static data publishing, if frequency of few sensitive values is very high in dataset as compared to other sensitive values then adversary can infer with high confidence the sensitive value of individual. We explain the problem by considering an example of a hospital that only deals with heart patients. Majority of patients of such hospital will have heart related disease or may be such type of disease which is due to heart problem. Let's consider that the microdata table have 3 distinct values {Angina, Diabetics, Tachycardia}. The microdata table 1 shows total 24 patients (including 21 Angina, 1 Diabetics, 2 High blood pressure).

Probability of any individual is very high to be heart patient. So some privacy model is required to decrease this probability to a level that adversary loses the advantage of data skewedness.

To resolve this problem, we are extending the approach of anatomy. We will divide the microdata table into 2 tables i.e. QIT and ST. In table 3a, we will keep the QI attributes values and will divide whole data into groups assigning them unique QID. In table 3b, we will have the QID from each group of table 1 and distinct sensitive attribute values of each QI group in first table. Table 3b, will have two additional columns which are distribution and weightage. Distribution column will have the percentage of each sensitive value in QI group respective to its frequency in whole table. Weightage column will have numeric value against each sensitive attribute in each QI group. Higher the weightage will indicate that frequency of particular sensitive value is higher in QI group.

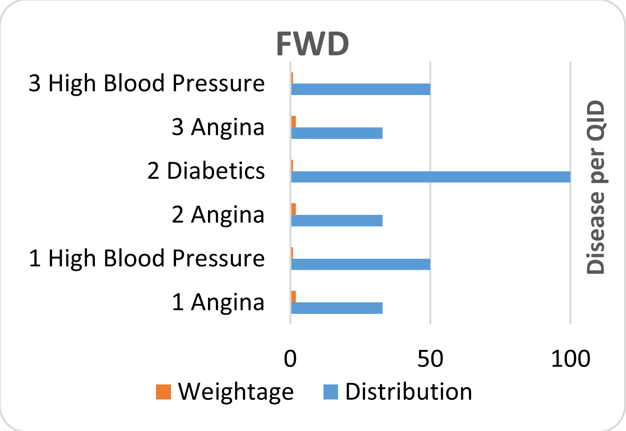

Figure 1: **FDW**

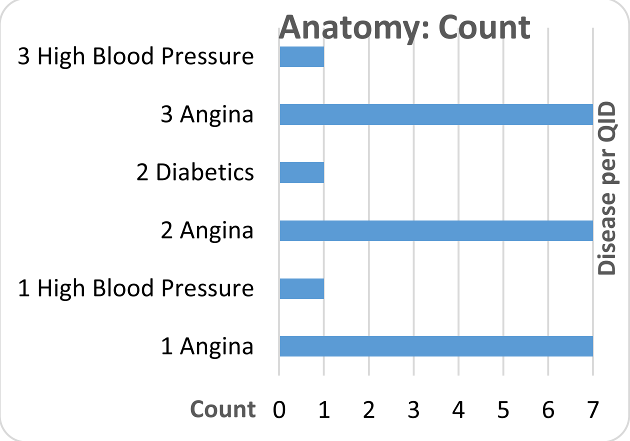

Figure 2: **Anatomy**

Age	Zipcode	Disease
35	2141	Angina
36	2141	High Blood Pressure
37	2138	Angina
38	2140	Angina
39	2140	Angina
40	2140	Angina
41	2140	Angina
42	2139	Angina
43	2141	Angina
44	2139	Angina
45	2140	Angina
46	2139	Angina
47	2140	Angina
48	2138	Angina
49	2139	Angina
50	2138	Diabetics
51	2139	Angina
52	2138	High Blood Pressure
53	2139	Angina
54	2138	Angina
55	2139	Angina
56	2139	Angina
57	2139	Angina
58	2138	Angina

Table 1: **The microdata**

QID	Disease	Count
1	Angina	7
1	High Blood Pressure	1
2	Angina	7
2	Diabetics	1
3	Angina	7
3	High Blood Pressure	1

Table 2: **The Anatomized (Sensitive) Table**

2.1. Comparison between FWD and Anatomy:

In figure 2, if we observe the results of anatomy in case of skewed data, then it is obvious that if patient is in QI group 1 then he is suffering from angina with high probability.

However, if we observe the figure 1, then it is more difficult for an adversary to make any guess by looking at distribution. The distribution of data is shown in figure 3.

Age	Zipcode	QID
35	2141	1
36	2141	1
37	2138	1
38	2140	1
39	2140	1
40	2140	1
41	2140	1
42	2139	1
43	2141	2
44	2139	2
45	2140	2
46	2139	2
47	2140	2
48	2138	2
49	2139	2
50	2138	2
51	2139	3
52	2138	3
53	2139	3
54	2138	3
55	2139	3
56	2139	3
57	2139	3
58	2138	3

Table 3a: **The quasi-identifier table (QIT)**

QID	Disease	Distribution	Weightage
1	Angina	33	2
1	High Blood Pressure	50	1
2	Angina	33	2
2	Diabetics	100	1
3	Angina	33	2
3	High Blood Pressure	50	1

Table 3b: **The sensitive table (ST)**

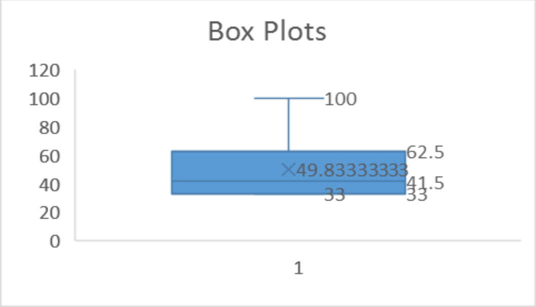

Figure 3: **Boxplot**

3. Conclusion:

In this paper, we have modified the anatomy approach to address the issue of data skewness. Instead of calculating the count, we will compute frequency distribution that shows the percentage of each sensitive value in QI group respective to its frequency in whole table. The data analyst can do effective data analysis by using the weightage table, to get frequency of particular sensitive value in each QI group.

Acknowledgement:

This work was supported by the Industrial Core Technology Development Program(10049079 , Development of Mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) *Dr. CS Hong is the corresponding author.

References

[1] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz., 10(6):571–588, 2002.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkita- subramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[3] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[4] Xiao, Xiaokui, and Yufei Tao. "Anatomy: Simple and effective privacy preservation." *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006.