Journal for Research | Volume 03| Issue 01 | March 2017
ISSN: 2395-7549

A Survey on Data Anonymization for Big Data Security

Athiramol. S
PG Student
Department of Computer Science & Engineering
St. Joseph's College of Engineering & Tech. Palai, India

Sarju. S
Assistant Professor
Department of Computer Science & Engineering
St. Joseph's College of Engineering & Tech. Palai, India

Abstract

Nowadays data analysis centers have a vital role in producing results that are beneficial for the society such as awareness about new disease outbreaks, the geographical areas affected by that disease, which aged people is mostly infected by that disease etc. The approach for protecting individual's privacy from attackers are well known as anonymization. The word anonymization in this context is hiding the information in such a way that illegitimate user should not be able to infer anything while legitimate user such as an analyzer should get sufficient information from it. That is the anonymization is stated in terms of security and information loss. There are different techniques used for anonymization. In this review, different anonymization techniques and their disadvantages are discussed. The main motto of all such anonymization is low information loss and better security. Although providing 100 percent security and 100 percent data utility is not possible for any systems as anyone of them compromises accordingly. All the techniques are based on concepts.

Keywords: Anonymization, Cryptography, l-Diversity, Multi Set based Generalization, Semantic Anonymization, Taxonomy Tree

I. INTRODUCTION

There are number of diseases that are reporting every year. From where we are getting an analysis on it?. It is done by various analysis centers around the country. Analysis centers gathers as much data as possible in order to perform analysis. In foreign countries, they are publishing data publicly such that it reaches ordinary people. There is a chance to get it misused any way. Various studies are being carried out based on this and found anonymization as a better option for preserving privacy and data utility. Anonymization is not performed on all the attributes present in the data being published. It is done only on the Quasi Identifiers (QID's). The QID's are nothing but the attributes which when as single may not disclose any information regarding an individual, but when used combined, will disclose the information. The anonymization is done in such a way that the adversary gets confused on seeing the equivalence classes that are generated as the result of anonymization.

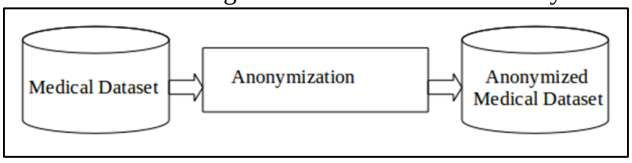

Fig. 1: A Privacy Model

The dataset which we want to anonymise is fed into an anonymization module, where the quasi identifiers are anonymized. A simple privacy model is depicted as shown in Fig. 1. The dataset thus produced termed as the anonymized dataset is allowed to get published. The major challenge in anonymising the dataset using any anonymization technique is that the information that can be inferred from that set should be useful in the same way privacy of every individual is preserved. There are different techniques that can be used for anonymising the dataset. K-Anonymity, L-Diversity, Cryptography, Taxonomy tree, Multi set based generalization, Semantic anonymisation, Scalable two phase specialization are some of them. Each of these techniques are having advantages and disadvantages. The data set consist of numerical as well as categorical attributes. The numerical attributes are suppressed so as to keep the anonymization level. The important matrices to keep in mind are the information loss (Suppression ratio) and disclosure risk. Suppression ratio is defined as the ratio between the numbers of suppressed tuples to the total number of records. The disclosure risk is evaluated as the ratio to number of tuples that can be identified individually to the total number of records. The solution that yields an optimal balance between these two are said to be a good anonymization algorithm. The possible attacks on the anonymized data are Homogeneity attack, Background knowledge attack, Probability inference attack. Homogeneity attack is kind of attack that can occur because of the ignorance to sensitive attribute from getting anonymized. Although the QID's are made identical for making the adversary confused, the sensitive attributes may have the same values thus making the anonymization process ineffective. Background knowledge attack is nothing but the knowledge of the adversary about an individual. This knowledge helps the attacker to narrow down possible values of the sensitive field further. Probability inference attack is performed based on the distribution of sensitive attribute values in an equivalence class.

II. LITERATURE SURVEY

A. K- Anonymity

L. Sweeney proposes a method called K-Anonymity where the data that are being published for analysis are anonymized in such a way that there will be atleast k individuals with the same data entries such that a particular individual will not get identified by a third party. The disclosure risk or the degree of privacy is directly proportional to the value of k. Greater the k value lesser will be the chance for attacks but greater will be the information loss. So k value should be less than a threshold level.

Table - 1
Sample Medical Data

Age	*Gender*	*Zipcode*	*Disease*
22	*M*	*12103*	*Headache*
25	*F*	*12104*	*Headache*
22	*M*	*12104*	*Headache*
25	*F*	*12103*	*Cough*

Table - 2
A 4- Anonymous Table

Age	*Gender*	*Zipcode*	*Disease*
*2**	*	*1210**	*Headache*
*2**	*	*1210**	*Headache*
*2**	*	*1210**	*Headache*
*2**	*	*1210**	*Cough*

K- Anonymity with k=4 applied to a medical data as shown in Table- 1 results in Table- 2. The table is anonymized in such a way that there will be at least k (4) members will be present in one equivalence class. K- Anonymity do not bother on the distribution of sensitive attribute values. The attacker predicts an individual with a probability 1/k.

B. L- Diversity

Ashwin Machanavajjhala, Johannes Gehrke et.al. Propose a method called L- Diversity which is a slight modification to K-Anonymity. It ensures that the sensitive attribute takes diverse values within the anonymity groups. It is introduced in such a way that the Homogeneity attack can be reduced. Table- 2 is an L- Diverse table with L=2. Here the equivalence class have 2 sensitive attribute values (Headache and Cough). From the table one can conclude that an individual is having Headache with a probability 75%, and if the adversary knows that an individual have low risk for Cough, then the attacker can infer easily that a person is having Headache.

C. Cryptography

As there is a large amount of toolset for cryptography and its model is well defined, it is used widely for data mining. According to new studies, it is a proven fact that the cryptography techniques are able to reduce the privacy leaks in the time of computation but it is unable to protect the result of computation. Due to this reason there is a steep reduction in the usage of cryptography in the field of big data security.

D. Semantic Anonymization

Ahmed Ali Mubark, Hatem Abdulkader propose Semantic anonymization. It ensures the sensitive attribute values within an anonymity group is diverse semantically. For that some rules are defined in order to find a semantic relationship between two sensitive values.

Table - 3
A Table without Semantic Anonymization

Age	*Gender*	*Zipcode*	*Disease*
*2**	*	*1210**	*Blood Cancer*
*2**	*	*1210**	*Leukemia*

In Table- 3, one individual is having Leukemia and the other is having Blood cancer. They are grouped together although the two sensitive values are semantically similar. The semantic anonymization ensures that no two individuals with the semantically similar attributes come together in one group.

E. Multi-Set Based Generalization

Li, Tiancheng, et al. Introduces multi set generalization approach. It preserves the exact values of each attribute without being suppressed. In Table- 1, each attribute is having a single value. Privacy can be easily breached as no security measures are adapted.

After performing the multi set based anonymization, the exact values are not only free from direct inferences but also preserves their occurrences. Table- 4 shows the generalized table.

Table - 4
Multi set based Generalized Table

Age	*Gender*	*Zipcode*	*Disease*

22:1, 25:2	*M:2, F:2*	*12103:2, 12104:2*	*Headache*
22:1, 25:2	*M:2, F:2*	*12103:2, 12104:2*	*Headache*
22:1, 25:2	*M:2, F:2*	*12103:2, 12104:2*	*Headache*
22:1, 25:2	*M:2, F:2*	*12103:2, 12104:2*	*Cough*

The problem with this technique is that, the privacy within a bucket can be breached easily.

F. Scalable Two-Phase Top down Specialization (TPTDS)

Xuyun Zhang, Laurence T. Yang, propose Scalable two- phase top down specialization approach. It make use of Hadoop MapReduce so as to reduce the execution time. The task of performing anonymization is split into different MapReduce tasks and are performing it using multi node environment. As per the paper they could reduce the execution time but they have used K-anonymity for the anonymization of the records. So there will be chances of homogeneity attack as well as background knowledge attack to occur. TPTDS partitions the entire data before applying the specialization. The generalization is performed from the top most node of a taxonomy tree. Each specialization is decided by an information metric called IGPL value.

G. Taxonomy Tree Method

As the main aim of anonymization is individual's security with minimal information loss, this method is a best one. In this approach the categorical attributes are generalized according to the taxonomy tree for each of the attributes. As the sensitive attribute values gets changed from specific values to generalized value, the adversary will be unable to narrow down the possible values of sensitive attributes, thereby decreasing the chances for background knowledge attack. This technique is performed with an assumption that generalized is secure than specialized. Fig. 2 shows an illustration of taxonomy tree for disease attribute.

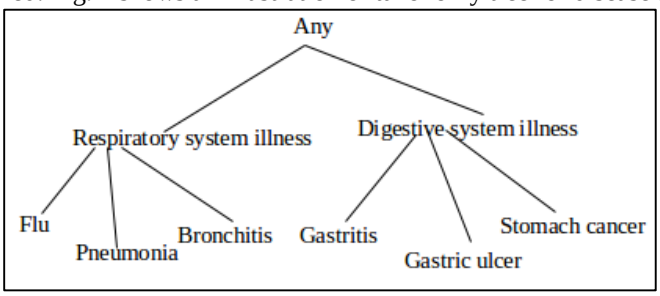

Fig. 2: Taxonomy Tree

III. COMPARISON CHART

Table – 5
Comparison of Various Anonymization Techniques

Method	*Deals*	*Ignores*
Cryptography	*Privacy leaks in the process of computation*	*Scalability*
K- Anonymity	*Direct inference attack*	*Homogeneity attack, Background knowledge attack*
L- Diversity	*Homogeneity attack*	*Background knowledge attack, Probabilistic inference attack*
Multi set based generalization	*Information loss, Direct inference attack*	*Background knowledge attack, Probabilistic attack*
Semantic anonymization	*Semantic relationship among sensitive attribute values within a group*	*Background knowledge attack, Probabilistic inference attack*
Scalable two phase specialization	*Execution time, Scalability*	*Background knowledge attack, Homogeneity attack*
Taxonomy tree	*Background knowledge attack, Information loss*	*Execution time*

IV. CONCLUSION

From the study it is clear that as there exist so many techniques in the field of anonymization, and it is a hot research topic nowadays. They all have a common drawback which is the background knowledge attack. As we are not able to predict the level of background knowledge an attacker has about an individual, we need to compromise slightly with the information loss. In that way, if we could generalize the sensitive attribute also it can reduce the background knowledge attack. It takes more time for execution. That can be reduced if we are using a Hadoop MapReduce execution model.

REFERENCES

[1] L. Sweeney (2002, May.). K-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, [On line]. 10(5), pp. 557-570. Available:
[2] https://epic.org/privacy/reidentification/Sweeney_Article.pdf
[3] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren (2014, Oct). Information security in big data: Privacy and data mining, pp. 1149-1176. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6919256
[4] Benny Pinkas , "Cryptographic techniques for privacy preserving data", SIGKDD Explorations, 4(2), pp. 12-19
[5] Li, Tiancheng, et.al. "Slicing: A new approach for privacy preserving data publishing," IEEE Transactions Knowledge and Data Engineering, pp. 561- 574, 2012.
[6] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, "l-Diversity: Privacy beyond k- anonymity," 2005.
[7] Xuyun Zhang, Laurence T.Yang, Chang Liu, and Jinjun Chen, " A scalable top down specialization approach for data anonymization using MapReduce on cloud", IEEE Transactions on parallel and distributed systems, pp. 363-373, 2014
[8] Ahmed Ali Mubark, Hatem Abdulkader, "Semantic anonymization in publishing categorical sensitive attributes, IEEE Int. Conf, 12;pp. 89-95, 2016.