International Journal of Electronic and Electrical Engineering.
ISSN 0974-2174 Volume 7, Number 8 (2014), pp. 773-778
© International Research Publication House
http://www.irphouse.com

Comparative Analysis of Anonymization Techniques

Dilpreet Kaur Arora[1], Divya Bansal[2] and Sanjeev Sofat[3]

*[1, 2, 3]Computer Science Department,
PEC University of Technology, Chandigarh, India*

Abstract

In recent years, privacy-preserving techniques has seen quick advancement due to rapid increase in storing and maintaining personal data about individuals. The personal data can be misused, for a variety of purposes. Maintaining the privacy for high dimensional database has become major aspect. In order to improve these concerns, a number of Anonymization techniques have recently been proposed in order to perform privacy-preservation of data. In this paper, a comparative analysis for K-Anonymity, L-Diversity and T-Closeness Anonymization techniques is presented for the high dimensional databases based upon the privacy metric.

Keywords Anonymization, K-anonymity, L-diversity, t-closeness, Attributes.

Introduction

Due to the rapid growth in information technologies, companies at the present time collect and store huge amounts of information in their databases. Typically, such information is stored in the form of tables and each record is corresponding to an individual. Every record has a number of attributes which can be divided into three categories: 1. Explicit identifiers which can clearly identify individuals. 2. Quasi Identifying attributes whose values when taken can easily identify individuals identities. 3. Sensitive Attributes which are considered sensitive and need not be disclosed[4].

A number of different Anonymization techniques have been researched to protect the identity of the respondents. Different data holders like often remove or encrypt the explicit identifiers. While de-identifying the information which does not provide anonymity, as released information also contains other data called Quasi Identifiers which can be used for re-identifying the data respondents, thus leaking that information which is not intended to be disclosed. While releasing the information, it is necessary to protect the sensitive information of the individuals from being disclosed. While the released table gives useful information to the researchers, it also presents disclosure risk to the individuals whose data is present in the table. Therefore, the major objective is to limit the disclosure risk to an adequate level. This can be achieved by anonymizing the data before releasing it. To efficaciously limit disclosure, we need to evaluate the disclosure risk of an anonymized table. In this paper, we are performing a comparative analysis of Anonymization techniques on the basis of privacy and performance.

The rest of this paper is structured as follows: Section 2 describes classification of Anonymization techniques; followed by comparative analysis of those techniques which have been briefed in Section 3; finally Section 4 concludes the review.

Classification of Anonymization Techniques:

Data Anonymization is the process applied on the data to prevent identification of individuals, making it possible to share and analyze data securely[11]. Figure 1 shows the classification of different Anonymization techniques and the algorithms used by those techniques.

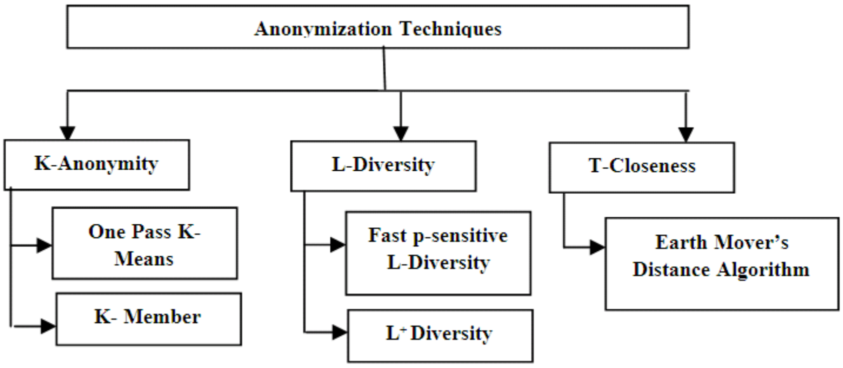

Figure 1: Classification of Anonymization Techniques

K-Anonymity

Sweeney [1] introduced k-anonymity as the property that each record is indistinguishable with atleast k-1 other records with respect to the quasi-identifier. A common Anonymization approach i.e. generalization at cell level was introduced.

According to Jun-Lin Lin et. al. [3] an efficient clustering method for K-Anonymization is used. The algorithm used is One Pass K-Means Algorithm which is being divided into two stages: Clustering Stage and Adjustment Stage.

In this paper Ji-Won Byun et.al [5] proposed an approach that uses the idea of clustering to minimize information loss and thus ensure good data quality. In this paper, Greedy K-Member clustering Algorithm is used for investigating the performance in terms of data quality, efficiency and scalability.

Incognito: Efficient Full-Domain K-Anonymity proposed by Raghu Ramakrishnan et.al[13] provides a practical framework for implementing full domain generalization and categorization of taxonomies of K-Anonymization techniques.

B.K.Tripathy et.al [10] proposed Kernel Based K-Means Clustering Using Rough Set which is a nonlinear transformation. In this paper different clustering techniques are being considered one of the first algorithms deals with uncertainty in fuzzy K-means [11]. Krishnapuram and Keller [12] proposed a probabilistic approach to clustering.

According to V. Ciriani et.al [9] K-Anonymity is broken down into two approaches i.e. generalization and suppression. Xiaokui Xiao et.al [8] proposed the concept of personalized anonymity which performs minimum generalization and retains a large amount of data.

L-Diversity

An equivalence class is said to have L-diversity if there are at least L "well-represented" values for the sensitive attribute. A table is said to have L-diversity if every equivalence class of the table has L-diversity [4].

Ashwin Machanavajjhala et.al [6] noticed two attacks which can take place in K-Anonymity. The author's proposed a novel definition called L-diversity which states that "A q-block is L-diverse if contains at least L "well-represented" values for the sensitive attribute S.

B.K.Tripathy et.al [2] proposed a fast p-sensitive l-diversity Anonymization algorithm. In this paper, a third phase named L-Diversity Algorithm is added to the two phases of OKA to achieve l-diversity.

T-Closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t[4]. The requirement of t-Closeness is the distribution of sensitive attributes in any eq. classes is close to the distribution of a sensitive attributes in the overall table [13].

T-Closeness: Privacy Beyond K-Anonymity and L-Diversity discussed by Ninghui Li et.al [4] proposed Earth Mover's Distance which is used to calculate the distance between the above mentioned requirements.EMD is calculated for two Attributes: Numerical Attributes and Categorical Attributes.

Further Luo Yongcheng et.al [7], concluded that protecting data privacy is an important issue while collecting microdata. K-anonymity can resist the links attack, l-diversity can resist against the homogeneous attack [6], t-closeness will be able to minimize the loss information.

Comparative Analysis:

In this section we implemented these techniques for performing comparative analysis on the basis of a privacy metric i.e. Information loss. For privacy comparison several real world datasets were being used. The datasets includes transportation system dataset, census marriage dataset and Crime state by state dataset which are as follows:

1. Transportation Dataset (Size: 465): It Contains six attributes IMEI Number, Latitude, Longitude, X, Y, Z values of accelerometers and gyroscopes out of which IMEI Number is sensitive attribute and rest of them are quasi-identifying.
2. Marriage Dataset (Size: 2348): It Contains five attributes Year, Age, Marital Status, Gender and people out of which age is a sensitive attribute and rest of them are quasi-identifying.
3. State wise Crime Dataset (Size: 16422): It Contains five attributes State, Type of Crime, Crime, Year and count out of which state acts as sensitive attribute and rest are quasi-identifying.

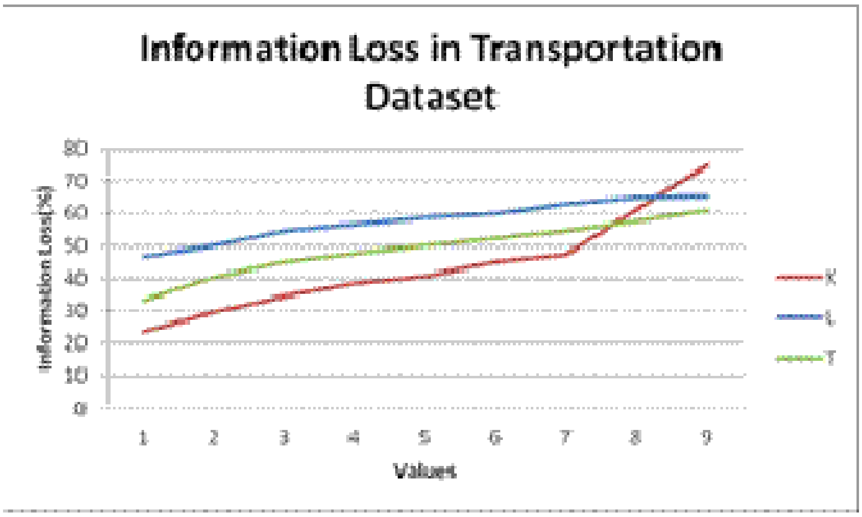

Figure 2: Information Loss in Transportation Dataset

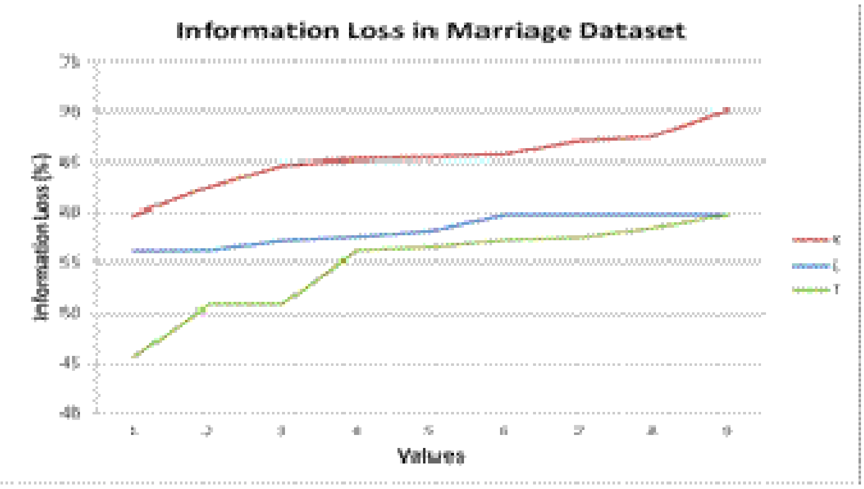

Figure 3: Information Loss in Marriage Dataset

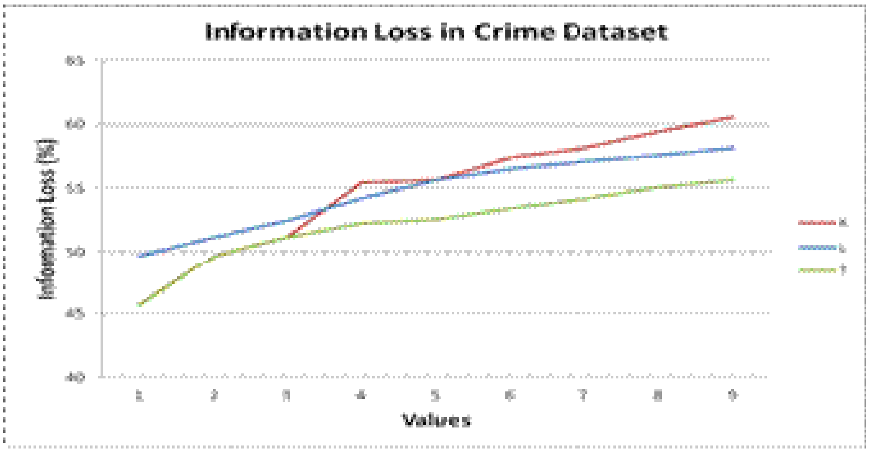

Figure 4: Information Loss in Crime Dataset

Above Figures depicts the comparison of different Anonymization techniques i.e. K-Anonymity, L-Diversity and T-Closeness for different datasets based upon a privacy metric Information Loss. Here the values of K, L and T denote the no. of attributes to be anonymized. As the no. of attributes increases the information loss gets increased showing that the information loss is directly proportional to the no. of attributes to be anonymized.

So from the above comparison we concluded that the T-closeness causes consistently less information loss than l-Diversity and K-Anonymity.

Conclusion and future Work

It became obvious from the literature that privacy of users is the major concern these days. Various models proposed for microdata have been adopted for preserving privacy of different types of data like transportation system data, medical data, marriage census data etc. This paper reviews the various Anonymization techniques like K-anonymity, L-diversity and T-closeness. These techniques are analyzed for different datasets and it is concluded that T-Closeness has less information loss than L-Diversity and K-Anonymity but these techniques still leads to extensive information loss. So, there is a scope of enhancement of the techniques that will provide privacy preservation with minimum information loss and better utility of released data. In future we would also compare these techniques on other metrics.

Acknowledgement

The authors would like to thank Deven, Abhiyanshu and Nidhi Undergraduate students of PEC University of Technology, Chandigarh for providing the Transportation data through the application developed by them.

References

[1] Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression, In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems Volume 10 Issue 5, 571-588 (2002)

[2] Tripathy, B., Maity, A., Ranajit, B., Chowdhuri, D.: A fast p-sensitive l-diversity Anonymisation algorithm. In: Recent Advances in Intelligent Computational Systems (RAICS), pp. 741-744. IEEE Press, Trivandrum (2011)

[3] Lin, J., Wei, M.: An Efficient Clustering Method for k-Anonymization. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society, pp. 26-35.ACM, New York (2008)

[4] Li, N., Li, T., Venkatasubramanian, S.: T-Closeness: Privacy Beyond k-Anonymity and L-Diversity. In 23rd IEEE International Conference on Data Engineering, pp. 106-115.IEEE Press, Istanbul (2007)

[5] Byun, J., Kamra, A., Bertino, E., Li, N.: Efficient k-Anonymization Using Clustering Techniques. In Kotagiri, R., Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) Advances in Databases: Concepts, Systems and Applications. LNCS, vol.4443, pp.188-200. Springer, Heidelberg (2007)

[6] Machanavajjhala, A., Gehrke, J., Kifer, D.: L-Diversity: Privacy Beyond K-Anonymity. In 22nd International Conference on Data Engineering, pp. 24. IEEE Press, Atlanta, GA, USA (2006)

[7] Yongcheng, L., Jiajin, L., Jian, W.: Survey of Anonymity Techniques for Privacy Preserving. In International Symposium on Computing, Communication, and Control, pp.248-252. IACSIT Press, Singapore (2009)

[8] Xiao, X., Tao, Y.: Personalized Privacy Preservation. In international conference on Management of data, pp. 229-240.ACM, New York (2006)

[9] Ciriani, V., Vimercati, V., Foresti, S., Samarati, P.: k-Anonymity. In Kikuchi, Hiroaki, Rannenberg, Kai (eds.) Advances in Information and Computer Security. LNCS, vol.4752. Springer, Heidelberg (2007)

[10] Tripathy, B., Ghosh, A., Panda, G.: Kernel Based K-Means Clustering Using Rough Set. In International Conference on Computer Communication and Informatics, pp. 1-5.IEEE Press, Coimbatore (2012)

[11] Data Anonymization, http: //www.slideshare.net/KaiX/ lions-zebras-and-big-data-anonymization

[12] l-diversity and t-closeness, http: //www.utdallas.edu/~muratk/courses/ dbsec12f_files/DBSec_priv3.pdf

[13] LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: Effcient Full-Domain K-Anonymity. In international conference on Management of data, pp.49-60.ACM, New Yorsk (2005).