Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 2 (2017) pp. 247-253
© Research India Publications
https://dx.doi.org/10.37622/ACST/10.2.2017.247-253

Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data

Johny Antony P
Research Scholar
NGM College, Pollachi, Tamilnadu, India.

Dr. Antony Selvadoss Thanamani
Head, Department of Computer Science
NGM College, Pollachi, Tamilnadu, India.

Abstract

Modern technology and networking generates huge volume of data . Privacy of data is a crucial issue and a topic for significant research. Data publishing faces the problem of deciding how to publish useful data while preserving privacy-sensitive information according to the privacy requirements of data holders. According to the concept of the privacy protection, it is defined as such the accessing of published data must not allow the unwanted users to identify anything about the targeted individuals. This paper presents a classification and analysis of various anonymization techniques for privacy preservation like k-anonymity, l-diversity, t-closeness, differential privacy, slicing.

Keywords: k-anonymity, l-diversity, t-closeness, differential privacy, slicing, data utility

1. INTRODUCTION

Present information technology creates vast amount of data characterized by velocity, volume, veracity. Sharing and dissemination of this data gives rise to the violation of privacy of individuals who are the subjects of the data. Privacy protection is one of most important issue in big data processing. Customer privacy is an issue that attracts considerable attention from both academia and IT industry [2]. The question that comes to the mind is that can share the data while protecting the privacy and at the same time providing the data utility. It opens new avenues for research and study as privacy is a right of an individual. The primary goal is to extract the hidden wisdom and knowledge from the huge amount of data at the same time sensitive data should not be misused. Despite the use of big data for innovation and insights, the massive amount of data can breach the privacy of users. In order to preserve privacy of data several mechanisms have been proposed and developed in the recent years. It is a great challenge to keep balance between data utility and data privacy. [3]

This paper presents the basic models of privacy preservation, their comparative study and performance with regard to execution time, data utility and privacy preservation.

2. RELATED WORK

Sweeney presents k-anonymity as a model for protecting privacy. k-anonymity is one of the basic privacy preservation model [17]. Machanavajjhala, A et.al. proposed l-diversity. [6]. Li et.al. presents t-closeness as a basic model for privacy preservation [5]. They propose this model as beyond k-anonymity and l-diversity. Sapana Anant Patil and Dr. Abhijit Banubakod [14] made comparative study of privacy preserving techniques in data publishing . Ram Mohan Rao P [12] made comparative study of privacy preservation techniques in data analytics. M. Nithya and Dr. T. Sheela [7] studied on privcy preserving data mining techniques. Priyank Jain et.al [11] made a comparison on privacy preservation methods for big data.

3. BASIC MODEL IN PRIVACY PRESERVATION

3.1. k-anonymity

k-anonymity is one of the basic privacy preservation model. In the k-anonymity, every published record has to be indistinguishable from at least (k-1) others on its QI attribute. The "quasi-identifiers" are the attributes available to an adversary. It is defined as: A table T satisfies k-anonymity if for every tuple $t \in T$ there exist $k-1$ other tuples $t_{i1}, t_{i2}, \ldots, t_{ik-1} \in T$ such that $t[C] = t_{i1}[C] = t_{i2}[C] = \ldots. = t_{ik-1}[C]$ for all $C \in Q$. [17].

3.2. l-diversity

l-diversity is a group based anonymization model that assists to preserve the privacy of data through reducing the granularity of a data representation using generalization and suppression. In l-diversity, an equivalence class is said to have l-diversity if there is at least l “well-represented” value for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. It is defined as: A q block is ℓ-diverse if it contains at least ℓ “well-represented” values for the sensitive attribute S. A table is ℓ-diverse if every q block is at least ℓ-diverse [6,3].

3.3. t-closeness

t-closeness is another group based privacy model that extends the l-diversity model. It treats the values of an attribute distinctly, and considers the distribution of data values of the attribute to preserve the privacy. It uses the Earth Mover Distance (EMD) function to compute the closeness between two distributions of sensitive values. It is defined as: 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 not more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness [5].

3.4. Differential Privacy

Differential Privacy offers one way forward that to extract insights from a database while guaranteeing that no individual can be identified. It achieves the guarantee of privacy by adding noise to answer to the queries. The amount of noise added must be large enough to conceal the effect of individuals and small enough that does not distort the genuineness of the answer.It is defined as : Let databases (D, D’) differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row. A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and all S ⊆Range(K) [7].

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D') \in S]$$

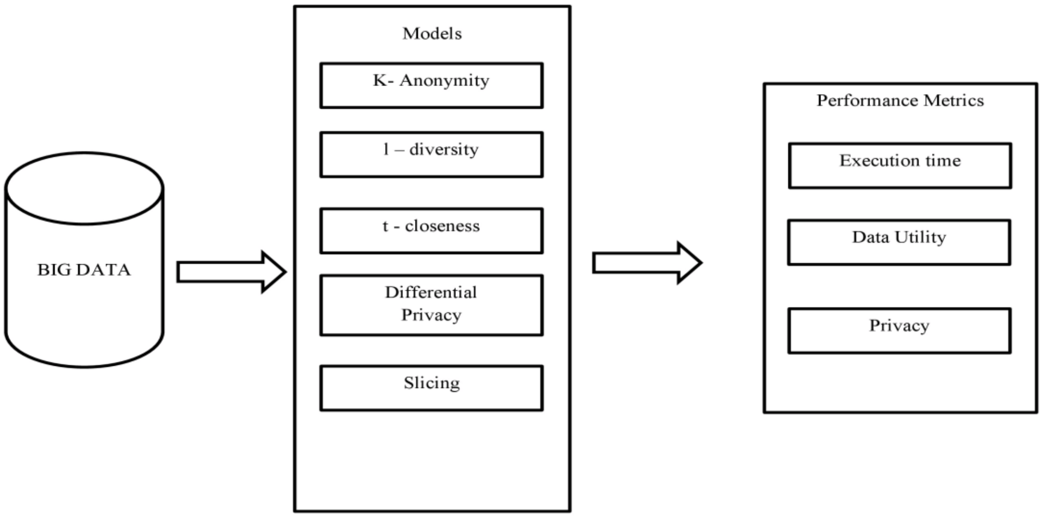
3.5.Slicing

It is a technique that partitions data horizontally and vertically. The basic idea is to break the association between cross columns but to preserve the association within each column. Slicing preserves data more accurately. [14].

Table 2. Comparison of Privacy Models

No	Model	Merits	Demerits
1	Randomization	Simple method that can be easily implemented	Difficulty for multiple attributes and categorical attributes
2	k-anonymity	Easy to implement, Chance for Re-identification is less when the value of k is high.	It fails in preventing the background knowledge and homogeneity attacks, Suffers from attribute linkage and record linkage, Long processing time, Utility may be compromised that any query returns minimum of k matches.
3	l-diversity	Reduce the data set into summary form. Sensitive attribute would have at most same frequency.	Depends upon the range of sensitive attributes. For l diverse, there should be l different values of sensitive attribute. It is prone to skewness and similarity attack and may not prevent

			attribute disclosure. Vulnerable to homogeneity attach and back ground knowledge attack.
4	t-closeness	Prevent data from skewness attack.	Complex computational procedure to enforce t-closeness. It looses the co relation between different attributes since each attribute is generalized separately. Utility is damaged when t is very small
5	Differential Privacy	Most suitable for big data. Provides strongest privacy guarantee.	Data utility may be reduced. Data miner is only allowed to pose aggregate queries. Probability of attacking both the databases by adversary is not taken into consideration.
6	Slicing	Randomization on sensitive attributes. Prevents attribute disclosure.	Utility and risk measure is not matched. It may break association between attributes.



Architecture

4. EXPERIMENTAL RESULTS

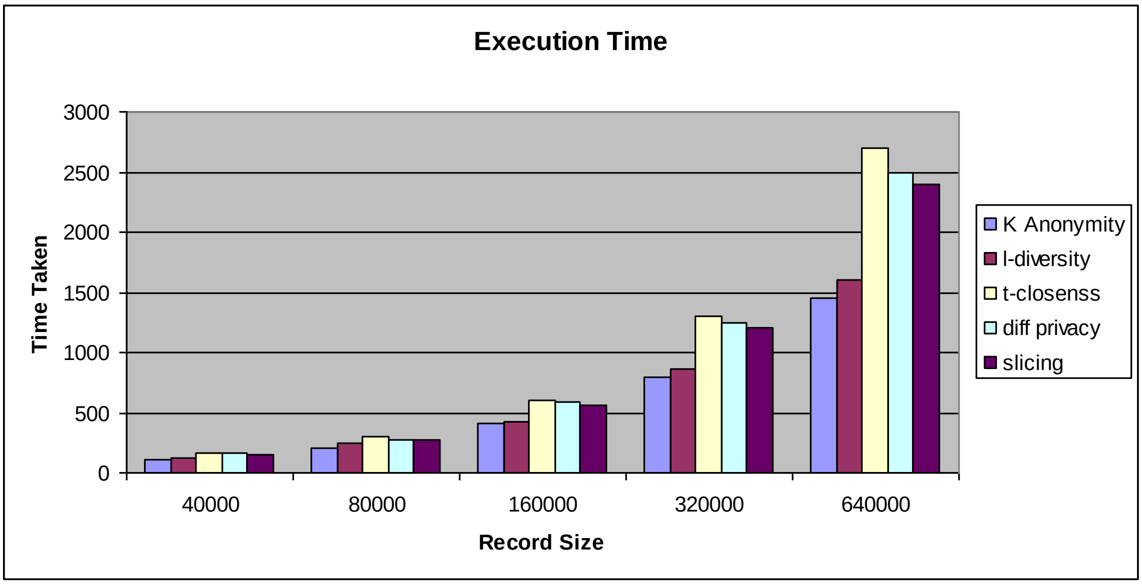
Experiments were carried out on adult data set taken form UC Irvine Machine Learning Repository - UCI Machine Learning. (https://archive.ics.uci.edu/ml/datasets.html). The data set contains 48842 instances with 14 attributes both categorical and integer. The data contains sensitive and non sensitive (quasi identifier) attributes. The data was cleansed and formatted and made into sets of 40000, 80000, 160000, 320000 and 640000 with random replication. The experiments are conducted on a machine with Intel ® Core TM i5-2120 CPU @ 3.30 GHZ, 4 GB RAM, Window 7, JAVA –JDK 8.0.

The objective of the experiment is to find out performance metrics such as execution time, data utility and privacy of the various privacy preservation models applied to big data.

4.1. Execution Time

The following table shows the execution time – the time taken by the algorithm to perform the task by various models with different data size.

Sl. No	Models/ Data Size	40000	80000	160000	320000	640000
1	k-anonymity	105	200	410	800	1450
2	l-diversity	130	240	430	860	1600
3	t-closeness	170	300	600	1300	2700
4	Differential privacy	160	280	590	1250	2500
5	Slicing	150	275	560	1200	2400



4.2. Data Utility & Complexity

Data utility is measured by the accuracy of the queries MIN, MAX, COUNT on the original data and the transformed data after applying the privacy preserving techniques.

Sl. No	Models	Data Utility	Complexity
1	k-anonymity	low	Very Low
2	l-diversity	high	Low
3	t-closeness	high	Very high
4	Differential privacy	medium	high
5	Slicing	medium	high

5. CONCLUSION

This paper gives a view about the basic model of privacy preservation and its effect when applied to big data. It also present the merits and demerits of each model for preserving privacy in data. An experimental result is also given in relation to execution time, implementation complexity and data utility.

REFERENCE

[1] Bayardo, R. J. and Agrawal, R., “Data privacy through optimal k-anonymization”, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), pp.217–228, 2005.

[2] Johny Antony P & Selvadoss Thanamani Antony., “A Survey on Privacy Preservation in Big Data”, International Journal of Engineering Science Invention Research and Development (IJESIRD) Vol 3, Issue 3, October 2016, ISSN 2349-6185

[3] Johny Antony P & Selvadoss Thanamani Antony., “A Review on Privacy Preservation in Big Data”, International Journal of Modern Computer Science and Application (IJMCSA) Vol. 4, Issue 6, November 2016, ISSN 2321-2632.

[4] Johny Antony P & Selvadoss Thanamani Antony., “A Privacy Preservation Framework for Big data using differential privacy and overlapped slicing”, International Conference on Big data infrastructure and cloud computing, Thiruvananthapuram, 9th October 2016, ISBN 9788192958040.

[5] Li, N., Li, T., and Venkatasubramanian, S., “t-closeness: Privacy beyond k-anonymity and l-diversity”, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 2006.

[6] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M., “l-diversity: Privacy beyond k-anonymity”, ACM Trans. Knowl. Discov. Data, Vol.1, No.1, 2007

[7] M. Nithya and Dr. T. Sheela., [2014], A Comparative Study on Privacy Preserving Data Mining Techniques, International Journal of Modern Engineering Research (IJMER), Vol 4, Issue 7, July 2014, ISSN 2249

[8] Mohammed, N., Chen, R., Fung, B., & Yu, P. S., “Differentially private data release for data mining”, In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493-501, ACM, 2011

[9] Mogre Neha V., & Patil Sulbha. (2013). Slicing: An Approach for Privacy Preservation in High-Dimensional Data Using Anonymization Technique, IRAJ International Conference, Pune 2013

[10] Patel Neha., Lade S., and Gupta R. (2015). Quasi and Senstive Attribute Based Perturbation Technique for Privacy Preservation. IJARCSSE vol 5, issue 11, November 2015

[11] Priyank Jain, Manasi Gyanchandani and Nilay Khare., Big data privacy: a technological perspective and review, Journal of Big Data, Springer 26 November 2016. DOI: 10.1186/s40537-016-0059-y

[12] Ram Mohan Rao P., [2016], Comparative study of Privacy Preservation in Data Analytics, International Journal of Innovation in Engineering and Technology, Vol 7, Issue 2, August 2016. IISN 2319-1058

[13] Rastogi Vibhor., Suciu Dan., & Hong Sungho (2007). The Boundary Between Privacy and Utility in Data Publishing. Journal of the ACM 978

[14] Sapana Anant Patil and Dr. Abhijit Banubakod., [2013], Comparative Analysis of Privacy Preserving Techniques in Distributed Database, International Journal of Science and Research (IJSR), Volume 4 Issue 1, January 2015 www.ijsr.net, ISSN (Online): 2319-7064.

[15] Sweeney Latanya Datafly: (2016). A system for providing Anonymity in Medical data, ACM

[16] Sweeney, l. (2002). Achieving k-anonymity Privacy Protection Using Generalization and Suppression,. International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems, 10(05), 571–588. doi:10.1142/s021848850200165x

[17] Sweeney, l. (2002). K-Anonymity: A Model For Protecting Privacy. International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems, 10(05), 557–570. doi:10.1142/s0218488502001648

[18] T.Malthi & S. Nandagopal (2014) Enhanced Slicing Technique for Improving Accuracy in Crowdsourcing Database, IJIRSET, Vol3, Issue1, February 2014