International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 9; Issue 06(E); June 2020; Page No.22625-22627
DOI: http://dx.doi.org/10.24327/ijcar.2020.22627.4470

Research Article

BIG DATA PRIVACY ANONYMIZATION ALGORITHMS: A REVIEW

Abhishek M and Prof. Poornima Kulkarni

Department of Information Science and Engineering RV College of Engineering Bengaluru, India

A R T I C L E I N F O

Article History:
Received 06th March, 2020
Received in revised form 14th April, 2020
Accepted 23rd May, 2020
Published online 28th June, 2020

Key words:

Anonymization, Privacy

A B S T R A C T

This paper provides an overview of big data privacy challenges, variety and classification of privacy algorithms. Information of each individual wish that his private information is not revealed in some or the other way. Privacy preservation plays a vital role in preventing individual private data. Anonymization is a process of removing personally identifiable information from the data set. Under anonymization category k-anonymity, L-diversity and T-closeness algorithms is discussed in detail. The comparison of anonymization algorithms is carried out for eight parameters.

INTRODUCTION

The big data is a huge amount of data in the size of terabytes which consists of structured data, semi-structured and un-structured data. Structured data is relational data which is in rows and columns. Unstructured data will be in the format of Json, xml. Semi-structured data is combination of both structured and unstructured data.

Big data can be well explained by its characteristics. The characteristics of big data are 3 V's volume, velocity and variety. Volume the name big data is related to a size which is enormous. Data can be considered as big data depending upon the volume of the data. Variety refers to heterogenous sources and nature of data. Earlier days spreadsheet and databases were the only sources of data. Now a days data in the formof audio, videos, mp3, emails are also considered in the data analysis process. Velocity refers to the speed of data generation and transfer to meet the demands.

Handling a large amount of data is challenging since there is huge increase in data (sources are media, Internet of things, web). The challenges of big data while processing, storing and fetching dataare:

1. Heterogeneity and incompleteness of available data.
2. Security and Privacy.
3. Storing and processing of data.
4. Visualization.

The most concerning issue in big data analytics is privacy and security of the personalized information. Since privacy and security is very essential for any organization or an individual, securing such a huge data set from inside as well as outside becomes one of the major challenging issues of big data. [6]

Corresponding author:* **Abhishek M
Department of Information Science and Engineering RV College of Engineering Bengaluru, India

Literature Survey

In reference [1] This paper gaveInsights of big data, characteristics, lifecycle, security and privacy. Overview of privacy algorithms like K-anonymity, L-diversity and T-closeness. The privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy preserving techniques are sufficient for big data processing.

In reference [2] this paper gave overview of the tools and methodologies for data privacy protection that can cope with the challenges. Challenges of big data anonymization is discussed. Big Data and Knowledge Discovery are possible trends for evolving systems, providing privacy protection through anonymization techniques can help improving the dependability of these systems.

In reference [3] Different kinds of differential algorithms have been discussed. Improvement of differential privacy is achieved by choosing the epsilon. There are still challenges that may demand for schemes to be used in combination with cryptography, which is an area that requires further researches.

In reference [4] discussion and comparison of anonymization algorithms like datafly and mondain which comes under K-anonymization. By comparison it is found that datafly algorithm is moresuitable for synthetic dataset while Mondrian algorithm is more suitable for real dataset and other thirteen parameters comparison is done.

In reference [5] this paper discussed advantages and disadvantages of various algorithms suggested by various authors in last decide. Most of the suggested algorithms

assumed a single data release from a publisher, thus only protected the data up to the first release or the first recipient.
In reference [6] this paper has reviewed different privacy preserving techniques available for Big Data with their advantages and disadvantages. The paper also presents Differential Privacy Preserving technique which the most suitable technique for Big Data as it hides individual information while preserving the privacy of the whole data set.

In reference [7] this paper discussed a framework is developed which puts an additional feature of access control with privacy mechanisms for multiple role. When the access control and anonymization algorithms for data are integrated, they work together as an administration for an application as a configurable privacy protection for access control framework.
In refence [8] this paper gave Overview of Privacy-Preserving Data Publishing and systematic performance evaluation, in terms of efficiency and data utility, of three of the most cited k-anonymization algorithms.

Privacy

Data security is commonly defined as data secrecy, transparency and honesty, meaning the data is not accessed by unauthorized access or not meant to be used. Privacy protection is largely essential to comply with laws and legislation (it is the duty of a corporation to safeguard its data).Preservation of privacy is closely linked to the principle of avoiding the release of information.[4]

Data privacy is the proper use of the data. When companies use data that is provided to them, the data should be used according to the agreed purposes without disclosing the sensitive information. Eg: In college, they will have all the details of student including the bank details, in the website only the student name and USN is disclosed here the bank account is sensitive information it is not disclosed to public. Failure to provide sensitive information with privacy is not ethical, and may result in illegal activity.

Generalization and Suppression

Before providing the generalization and suppression, it is important to remember that each record in a database has multiple attributes grouped into three groups. I primary core attributes (attributes that specifically distinguish persons, such as university seat number, Address , name, social security number); (ii) quasi-identifiers (attributes that can be paired with external information to reveal those people, such as date of birth, ZIP code, location, work, blood type);.(iii) sensitive attributes (attributes containing confidential individual details, e.g. bank account number, blood type, salary).

In the generalization technique the values of the attributes are generalized to a range to reduce the representation granularity. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. The attribute like age is stiffened into datasets. For example, age is generalized into age ranges. In suppression, quasi identifiers are replaced by some constant values like 0, *etc. The gender-like feature is removed from full dataset.

Anonymization

Data Anonymisation, also known as de-identification, consists of methods that may be used to discourage individual details from being retrieved. It is designed to reduce the information leakage about the individual although the data is shared and disclosed to the public. The method of anonymisation is done to alter the data before it is released.

Table 1 Patient data before anonymization

Sno	Zip	Age	Disease
1	57677	29	Cardiac problem
2	57602	22	Cardiac problem
3	57678	27	Cardiac problem
4	57905	43	Skin allergy
5	57909	52	Cardiac problem
6	57906	47	Cancer
7	57605	30	Cardiac problem
8	57673	36	Cancer
9	57607	32	Cancer

Table 2 After applying anonymization on Zip and age

Sno	Zip	Age	Disease
1	576**	2*	Cardiac problem
2	576**	2*	Cardiac problem
3	576**	2*	Cardiac problem
4	5790*	>40	Skin allergy
5	5790*	>40	Cardiac problem
6	5790*	>40	Cancer
7	576**	3*	Cardiac problem
8	576**	3*	Cancer
9	576**	3*	Cancer

classification

K-anonymity: It is a common and basic anonymisation technique approach. The aim is to reveal individual data without revealing sensitive attributes. K-anonymisation is achieved by methods of generalization and suppression. K-anonymity can be applied to data that has the principle of K-anonymity principle .A data is considered to have the value of K-anonymity if the details of each person can not be retried by at least k-1 individuals..If the K value is high, then the risk of disclosure to the data is low. Although it protects attributes of identifiers, it is vulnerable to attacks such as disclosure attack attributes and similarity attacks.

L-Diversity

L-diversity is the extension of k-anonymisation, and the limitations are overcome. In L-diversity the sensitive attributes within the data set must be diverse. When the overall distribution of sensitive attributes is skewed, L-diversity is not sufficient to protect privacy.L-diversity does not consider sensitive-value semantics. L·-diversity offers privacy preservation even in cases where the data publisher is not aware of the type of knowledge the rival requires to properly represent sensitive data in each group is L·-diversity's main objective. The system of L-diversity is prone to skew and similarity attacks and thus can not prohibit disclosure of attributes.

T-closeness

T-Closeness is a further improvement of the L-diversity method extended by the distinct treatment of an attribute 's values and the distribution of the attribute's data values to preserve privacy. It is said that an equivalence class has t-

closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the entire table is no more than a threshold t. A table is said to have t-closeness should be close to their distribution in the entire original database if all equivalence classes have t-closeness'-closeness distribution sensitive attributes within each quasi-identifier group.T-closeness using Earth Mover Distance (EMD) works to live the proximity of two critical value distributions, and requires the similarity to be at intervals. EMD is used, and is more effective, to quantify the distance between two distributions. Prevents disclosure attributes and attacks on skewedness. As the size and variety of data increases, so does the chances of data re-identification.

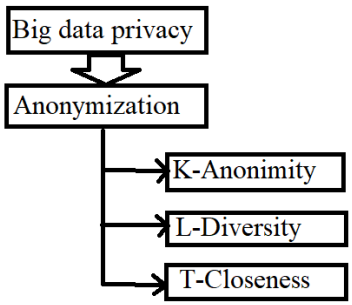

Fig 5.1 Classification of privacy algorithms

Comparison

Comparing anonymisation algorithms equally is necessarily a difficult job, since each proposed algorithm uses various settings and metrics. Algorithm performance may vary between various combinations of data sets and input parameters (e.g. an algorithm may work well in some experimental configurations and perform poorly in others). Anonymization algorithms are compared in detail based on 8 parameters.

Parameters	K-Anonymityzz	L-Diversity	T-Closeness
Execution time	Low	Low	High
Data utility	Low	High	High
Computation Complexity	O(klog k)	O(n^2)(K)	O(2^(nm))
No correlation among loss QID	NO	NO	YES
Attributedisclosure attack	Prone	Not prone	Not prone
Probabilistic attack	Not prone	Not prone	Not prone
Similarity attack	Prone	Not prone	Not prone
Skewness attack	Not prone	Prone(experiences)	Not prone

CONCLUSION

It has become very popular practice to exchange knowledge about the individuals. Maintaining the privacy of data publication is a very helpful method in this regard but maintaining the privacy of people and protecting confidential data is very important for every company.In this review, the study's main focus is privacy preservation using anonymization technique, and explaining a detailed study and comparison of three anonymization algorithms. We have done a thorough analysis of these three algorithms and obtained a comprehensive comparison based on eighth parameters of these two algorithms.

References

1. Priyank Jain, Manasi Gyanchandani and Nilay Khare, “Big data privacy: a technological perspective and review“ *Journal of Big data, Springer* open,2016.
2. Tania Basso, Roberta Matsunaga, Regina Moraes, Nuno Antunes, “Privacy Preservation In Big Data Using Anonymization Techniques “Seventh Latin-American Symposium on Dependable Computing,2016.
3. Xiaoming Yao, Xiaoyi Zhou, Jixin Ma, “Differential Privacy of Big Data: An Overview“ IEEE 2nd International Conference on Big Data Security on Cloud,2016.
4. Tanashri Karle, Prof. Deepali Vora, “Challenges on Anonymity, Privacy and Big Data“ International Conference on Data Management, Analytics and Innovation,2017
5. [5].Dr. Puneet Goswami, Ms. Suman ,Privacy Preserving “Data Publishing and Data Anonymization Approaches: A Review” Conference on Computing, Communication and Automation (ICCCA2017).
6. Ms. Sayyada Hajera Begum, Ms. Farha,”Nausheen A Comparative Analysis of Differential Privacy Vs other Privacy Mechanisms for Big Data”IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7.
7. Shankar Nayak Bhukya Dr.Suresh PabbojuDr. K Venkatesh Sharma ”Implementing Privacy Mechanisms for Data using Anonymization” Algorithms International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016
8. Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy”A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners” transactions on data privacy-2014.
9. Johny Antony P, Dr Antony Selvadoss Thanamani, “Differential Privacy Technique for Privacy Preservation on Big Data” *International Journal for Research in Applied Science & Engineering Technology* (IJRASET)-2019.
10. Priyank Jain, Manasi Gyanchandani and Nilay Khare, “Big data privacy: a technological perspective and review”springer open DOI 10.1186/s40537-016-0059-y.
11. Anjana Gosain andNikita Chugh “Privacy Preservation in Big Data” *International Journal of Computer Applications* (0975 – 8887)-2014.
12. P. Ram Mohan Rao1, S. Murali Krishna and A. P. Siva Kumar “Privacy preservation techniques in bigdata analytics: a survey”Journal of big data, Ram Mohan Rao *et al.* J Big Data (2018) 5:33https://doi.org/10.1186/s40537-018-0141-8.
13. Nancy Victor, Daphne Lopez and Jemal H. Abawajy ”Privacy models for big data: a survey” Int. J. Big Data Intelligence, Vol. 3, No. 1, 2016.