INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 6, Issue - 4, Apr – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 13/04/2020 Acceptance Date: 28/04/2020 Publication Date: 30/04/2020

K-Anonymization Technique for Privacy Preserving in Big Data

Kajol Patel
Lecturer, Computer Engineering,
Parul Institute of Engineering & Technology- Diploma Studies, Vadodara, India.
Email - Kajol.patel113@gmail.com

Abstract: *Big data has continues an rising in word of data analytics. Big data contain very large dataset and complex data structure. Traditional data model divided in the set of attributes like sensitive, quasi identifiers and non-sensitive attributes. Big data contain personal information that have privacy could be main. K-anonymization has been proposed as a mechanism for protecting privacy in big data. K-anonymization is a technique that prevents the above mentioned attacks by modifying the microdata which is released for business or research purposes. This is done by applying generalization and suppression techniques to the microdata. In this paper, k-anonymity is introduced and also some of the algorithms are studied which help in achieving k-anonymity.*

Key Words: *Big data, Data privacy, K-Anonymization, multidimensional.*

1. INTRODUCTION:

Many organizations are releasing microdata everyday for business and research purposes. This data does not include explicit indenters of an individual like name or address but it does contain data like date of birth, pin code, sex, marital-status etc. which when combined with other publicly released data like voter registration data can identify an individual. This joining attack can also be used to obtain sensitive information about an individual, thus, putting the privacy of an individual in grave danger.

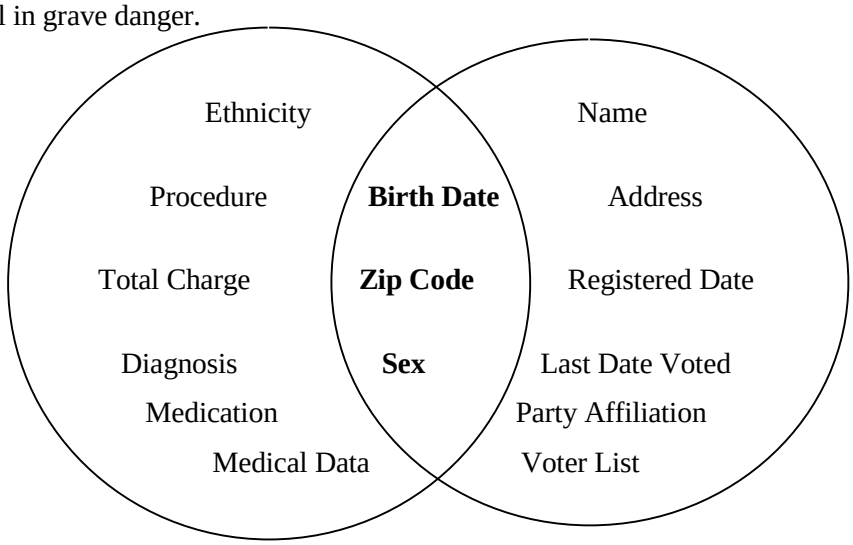

Fig. 1 Linking to re-identify data

Agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data is stored in a table, and each row corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:

- Attributes that clearly identify individuals. These are known as explicit identifiers and include Social Security Number, Address, and Name, and so on.
- Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip-code, Birthdate, and Gender.
- Attributes that are considered sensitive, such as Disease and Salary.

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in this paper: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity

disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm[17]. An observer of a released table may incorrectly perceive that an individual's sensitive attribute takes a particular value, and behave accordingly based on the perception. This can harm the individual, even if the perception is incorrect.

While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge or from other publicly-available databases that include both explicit identifiers and quasi-identifiers. A universal anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically reliable. Subsequently, more records will have a same position of quasi-identifier values. We characterize an equivalence class of an anonymized table to be a position of records that have the same values for the quasi-identifiers. As far possible revelation, we need to measure the disclosure risk of an anonymized table.

K-anonymity is one of the techniques which help us in release an enormous amount of data so that it very well used for business or research related work by various organizations by ensure that privacy of no individual is being placed in hazard because of the free data by defensive released information against inference and linking attacks. K-anonymity has the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protect against identity revelation, it is not enough to prevent attribute revelation.

In this paper, we have s tendency to describe varied measures, which will facilitate to make sure privacy in big data. This paper containing as follows: section II offers k- anonymization technique for big data. III We have tendency to discuss k- anonymization algorithm. Sections IV conclude our work.

2. METHODOLOGY:

K-anonymity is one of the techniques which help us in release an enormous amount of data so that it very well used for business or research related work by various organizations by ensure that privacy of no individual is being placed in hazard due to the released data by protecting released information against inference and linking attacks.

A. Basic Definitions

- **Key Attribute** - The attribute that can identify an individual directly is known as the key attribute. It is always removed during the release of data. E.g. Name, Mobile No.
- **Quasi-Indenter** - The set of attributes that can be used to identify an individual by using any means is called quasi-indenter. E.g. Date of Birth, PIN Code.
- **Sensitive Attribute** - The attribute containing the sensitive information about an individual is the sensitive attribute. E.g. Salary, Health Problem

B. K-anonymization

The K-anonymization is a framework for constructing and evaluating algorithm also system that release information. It is to allow sharing such data without compromising the privacy of the user. A data set called k-anonymiztion if for any k tuple with same quasi-identifier in the dataset there are at least k-1 other records that match those attributes[12]. K-anonymization also known as De-identification.

Name	Age	Gender	City	Disease
*	21<Age≤30	M	*	Cancer
*	21<Age≤30	M	*	Pneumonia
*	21<Age≤30	M	*	Dengue
*	21<Age≤30	M	*	TB
*	21<Age≤30	F	*	No illness
*	21<Age≤30	F	*	Viral infection
*	21<Age≤30	M	*	Heart related
*	21<Age≤30	F	*	Heart related
*	31<Age≤40	M	*	Viral infection
*	31<Age≤40	F	*	TB
*	31<Age≤40	F	*	Jaundice
*	31<Age≤40	M	*	Jaundice

Table 1 Base Dataset

Table 2 is a non anonymized database. There are four attributes along with twelve records in this data. There are two regular techniques for achieving k-anonymity for some value of k.

1. Suppression in this method, certain values of the attributes are replace by an asterisk '*'. In the anonymized Table 4 replaced all the values in the 'Name' attribute and each of the values in 'City' attribute by a '*'.

2. Generalization in this method, individual values of attribute is replaced with a broader category. For instance, the attribute 'Age' the value '24' replaced by broader category ' 21 <Age ⩽30.

Name	Age	Sex	City	Disease
Daksh	24	M	Delhi	Cancer
Jay	25	M	Gurgaon	Pneumonia
Vivek	28	M	Gurgaon	Dengue
Maulik	24	M	Delhi	TB
Kajal	26	F	Delhi	No illness
Shreya	27	F	Delhi	Viral infection
Neel	26	M	Delhi	Heart related
Krupa	30	F	Delhi	Heart related
Bhavin	32	M	Delhi	Viral infection
Shruti	39	F	Gurgaon	TB
Krishna	32	F	Gurgaon	Jaundice
Nirav	40	M	Delhi	Jaundice

Table 2 K-anonymized dataset

C. Models of k-anonymity

There are many possible combinations of different types of generalizations and suppressions result in different models of k-anonymity. The following are the different models:-

- **AG_TS:** Generalization is apply at the level of attribute (column) and suppression at the level of row.
- **AG_AS:** Both generalization and suppression are apply at the level of column. No specie approach has investigate this model. It should be noted that if attribute generalization is apply, attribute suppression is not needed. It becomes equivalent to AG.
- **AG_CS:** Generalization is apply at the level of column, while suppression at the level of set. It allows to decrease the consequence of suppression, at the price however of a higher complexity of the problem.
- **AG:** Generalization is applied at the level of column, suppression is not considered.
- **CG_CS:** Both generalization and suppression are applied at the cell level. Then, for a given attribute we can have values at different levels of generalization. By observations, this model is equivalent to CG.
- **CG:** Generalization is applied at the level of set, suppression is not considered.
- **TS:** Suppression is applied at the row level, generalization is not allowed.
- **AS:** Suppression is applied at the attribute level, generalization is not allowed. No explicit approach has investigated this model.
- **CS:** Suppression is applied at the set level, generalization is not allowed. Yet again it can be modeled as a reduction of AG.

3. CONCLUSION:

In this we described how micro-data released by various organizations for research or business purpose can compromise the security and privacy of an individual. In order to guarantee the anonymity of individuals and protect the released micro-data from any attacks we discussed the work that has been done in order to protect the released micro-data by the means of k-anonymity. Three basic algorithms for k-anonymity, under another name algorithm, Samarati's algorithm and Sweeney's algorithm, were studied upon where each of the algorithms had certain advantages and disadvantages. These algorithms were based on the AG model and AG_TS model of k-anonymity. We considered the effectiveness of these algorithms by plotting the graph between value of k and time taken to achieve k-anonymity with respect to the number of records. These algorithms have been really helpful in reducing the number of attacks on the micro-data and securing sensitive information of the individuals.

REFERENCES:

1. Priyank jain, Manasi Gyanchandani, Nilay Khare, "Big data privacy: a technological perspective and review", Sptinger,26 November2016.
2. Anjana Gosain, Nikita Chugh, "Privacy Preserving in Big data", International Journal of Computer Applications, August 2014.
3. Brijesh B.Mehta, udai Pratap Rao," Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges", ELSEVIER, volume 78, 2016,pp120-124
4. Fabian Prasser, Raffael Bild, Johanana Eicher, Helmut Spengler, Florian Kohlmayer, Klaus A. Kuhn, "Lighting: Utility-Driven Anonymization of High-Dimensional Data" , Transaction on data privacy 9,2016,Pages 161-185
5. Abid Mehomod, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, Song Guo," Protection of Big Data Privacy", IEEE Access, April 2016, Pages 1821-1834
6. J. Vinothkumar, V. Santhi," A Study on Privacy Preserving Methodologies in Big Data", IJST, December 2016, Pages 1-16
7. Reimann M, Schike O, Thomas J, "Toward an Understanding of Industry Commoditization", International Journal of Research in Markrting, 2010,pages 188-197
8. X. Zhang, C. Liu, S. Nepal,C.Yang,J. Chen, "Privacy Preservation over Big Data in Cloud System" , Security, Privacy and Trust in Cloud System, Springer, September 4, 2013,pages 239-257
9. J.Sedayo, "Enhancing cloud security using data anonymization", White paper, Inter Corporation.
10. Y.Gahi, M. Guennoun, Z. Guennoun, K. El-khatib, "Encryption Process for oblivious Data Retrieval", 6th International Conference for Internet Technology and Secured Transections,2011,pages 514-518
11. L. Sweeney, "K-anonymity: A model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge Based system,2002, pages 557-570
12. F.H.Cate, V.M.Schonberger, "Notice and Consent in a World of Big Data", Microsoft Global Privacy Summit Summary Report and Outcomes, November 2012
13. J. Salido, "Differential Privacy for everyone", White paper, Microsoft Coporation,2012
14. O. Heffets and K. Ligett, " Privacy and data-based research", NBER Working paper, September 2013
15. Cavoukain A. ,"Privacy by Scheme", Information and Privacy Commissioner of ontario, December 2016.
16. De Montjoye YA, Hidalgo CA, Verlesysen M, Blondel VD, "The Privacy bounds of human mobility", Scientific Report ,2013.
17. D. Lambert. Measures of disclosure risk and harm. J. Official Stat., 9:313,1993.
18. Latanya, "k-anonymity: a model for protecting privacy", International Journal of Uncertainty, Puzziness and Knowledge-Based Systems, Vol. 10, No. 5 (2002) 557-570