www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Volume-5, Issue-6, December 2017
Review Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

Comparison of various Anonymization Technique

G. Pannu[1*], S. Verma[2], U. Arora[3], A. K. Singh[4]

[1*] Department of Computer Applications, National Institute of Technology, Kurukshetra, India
[2] Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India
[3] Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India
[4] Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India

[*]*Corresponding Author: theshikhar3@gmail.com, Tel.: +91-8009-172-326*

Received: 10/Oct/2017, Revised: 24/Oct/2017, Accepted: 18/Nov/2017, Published: 31/Dec/2017

Abstract— Cloud based service is in trend for storing the database. Thus, exposing the data of the individual to the outside world is at the risk. Our major concern is to maintain privacy so that the data of the individual is not exposed to the adversary. In this paper, various techniques, how they have implemented, its new ideas and the models in order to implement privacy have been discussed. Few such techniques discussed are k-anonymity, l-diversity, t-closeness, (X, Y) anonymity, δ-Presence. All these techniques have its own approaches to secure data but in future, further new approaches having less time and space complexity can be thought of.

Keywords— Anonymization, Generalization, Supperession, Privacy

I. INTRODUCTION

Data mining is the process through which important information is retrieved from the large data sets [1]. In today's world the quantity of data is increasing day by day. One such example is Flickr, it is a public picture sharing website which approximately gets 1.8 million photographs per day. Just approximating the size of each photo to be 2 megabytes (MB), which means we require 3.6 terabytes storage space on every single day [2]. So, the above example gives us the idea of increasing size of database and the biggest challenge is to extract the knowledge and information from a huge volume of data. Data is collected in data collection phase where data holder collects data from the owner of the record and then the collected data is provided to the data miner during the publishing phase and then data mining is performed. Now a day, cloud services providers allow various health care institutes to store their database. As we know health care institutes are having information that is confidential and cannot be shared with any person. Though being a medical field there are phases where advice of various practioners are taken for further diagnosis and treatment. So, we have to decide which part of the database is to show so that confidently of the individual is not harmed and data can be sent for further diagnosis. Therefore, the data of the patient is at risk because it can be manipulated or misused by any person [3]. The need is to develop techniques such that the published data remains useful and privacy of the individual is also intact. Data privacy is the capability of organization to display only those fields of database to the third party that should not harm the privacy of the individual. This is also called privacy preserving data publishing. So, Anonymization is one such technique used to implement privacy preserving data publishing approach which helps to hide the identity of the record owner [4]. During privacy preserving data publishing, the table contains various entities such as Explicit identifiers, Sensitive attributes, Non-sensitive attributes, Quasi identifiers. Explicit identifiers are such attributes which explicitly identifies the individual. Sensitive attributes contain sensitive information of individual. Non-sensitive attributes are those which do not lie in the category of Explicit identifiers, Sensitive attributes and Quasi identifiers [5]. Adversary (attacker) can easily harm the privacy of individual, if he has the background knowledge. He can link various quasi identifiers to find the exact information of the person. Quasi identifiers are certain set of attributes which can be linked with some background knowledge to reveal the entire information of the individual. One such example: there was a governor of Massachusetts William Weld whose medical data was held by Group Insurance Commission. According to the voter list, six people were having the same birth date as of his and he was the one who is having five-digit ZIP code. This could be also possible through linkage attacks. If the adversary is able to link the records with the owner's record, to the data table, to the sensitive attributes then we call it as record linkage,

table linkage and attribute linkage respectively [6]. In all the three types adversary can find the record if he knew the QID of the individual. There were few approaches that has been proposed in order to preserve the privacy of the individual's record. Generalization use bottom-up strategy. In this, we replace values to the less specified values and operations performed on data set is based upon some specific range [7]. While suppression is replacing certain attributes with more specific data set [8]. Thus, in this technique the top-up approach is used and moves toward the less specified range then further these techniques are combined with the k-anonymity proposed by Samarati and Sweeney [9] where the QID value is same in at least k-1 records. Another method proposed by Machanavajjhala et al. [10] is l-diversity. It says there must be at least l "well represented "sensitive values. There is one such approach which improved version of k-anonymity that is (X, Y) anonymity proposed by Wang and Fung [11]. There is t-closeness model proposed by Li et al [12]. So, all these methods in their own way tried to maintain the privacy of the data. So, in this paper we tried to represent and describe various techniques that has been proposed and we have also compared their techniques and differentiated on various factors.

II. LITERATURE REVIEW

Before publishing the data, data publisher implements privacy model and when the data is safe, then it is published. Some of the attacks are Record linkage model, Attribute linkage model and Table linkage model. In the record linkage model some values 'x' in QID identifies the number of records in the table T. If the attacker is able to match the victim's QID with the value 'x' then the privacy of the victim is at risk. Attacker with some background knowledge can break the privacy of the person. To prevent record linkage techniques like k-anonymity and (X, Y)-Anonymity are used.

k-Anonymity

k-anonymity is a very basic model to implement the privacy on the provided dataset. K-anonymity says while performing the selection on the table for any combination of QID in the dataset, it should show at least k records in the result. In order to prevent record linkage, Samarati and Sweeney proposed a k-anonymity model. There must be at least k records having same QID. Thus, the minimum size of equivalence group should be k. So, we can say, the table satisfying the condition is k-anonymous. Thus 1/k will be at most probability of linking a victim to a specific record through QID [9,13].

Consider the following medical dataset with Name as explicit attribute and QID {Age, Job, Gender} and Disease as the sensitive attribute. K-anonymity is applied on the sample dataset of medical record as shown in Table I. On applying k-anonymity as shown in Table II, for any combination of QID{Age,Job,Gender} there should be at least k rows after performing selection on the records. So the above table shows two combination of QID[{35-40,professional,male},{30-35,Artist,female}] which is forming 3-k anonymity, because the number of records for QID {30-35,Artist,female} has the least number of records i.e 3. Suppose T is the table and |T| represents the number of records in the table. And 'x' is the search terms in the QID then $|\sigma x(T)|>=k$.

But this approach fails as there could be a situation, where the number of rows for a QID contains k rows and all rows contain same sensitive value in a column. So, the attacker can deduce the disease of the victim.

TABLE I. Sample dataset containing Medical data.

Name	Age	Job	Gender	Disease
Aman	35	Lawyer	Male	HIV
Komal	31	Singer	Female	Flu
Shubham	37	Teacher	Male	Cancer
Garvit	37	Engineer	Male	HIV
Manoj	38	Lawyer	Male	HIV
Kamal	36	Lawyer	Male	Cancer
Deepika	30	Singer	Female	Flu
Harshita	31	Dancer	Female	Cancer

TABLE II. 3-Anonymized Medical dataset.

Age	Job	Gender	Disease
35-40	Professional	Male	HIV
35-40	Professional	Male	HIV
35-40	Professional	Male	HIV
35-40	Professional	Male	Cancer
35-40	Professional	Male	Cancer
30-35	Artist	Female	Flu
30-35	Artist	Female	Flu
30-35	Artist	Female	Cancer

(X, Y)-Anonymity

To overcome the shortcoming of k-anonymity as discussed in table II, Wang and Fung [11] came up with the technique of (X, Y)-Anonymity where X and Y represents disjoint set of attributes in the table. Suppose in Table A, π(A) represents projection, σ(A) represents selection on A, |A| represents no of records in A, att(A) represents set of attributes in A.

Let i be the value of X. So, the anonymity of x with respect to Y is given by aY(x), which represents the number of distinct values on Y that will also occur in x. This can be represented as $|\pi Y \sigma x(A)|$. Also, a table satisfying (X, Y)-anonymity for any value of k will only be true if Ay(X)>=K.

(X, Y)-anonymity specifies that every value of X is linked to at least k distinct value on the column Y of the table A. Consider table I where X represents QID {Age, Job, Gender} and Y represents the sensitive value i.e. Disease

for the above medical dataset. Then (X, Y)-anonymity will represent that for every value of X in group, QID will be linked to diverse set of values in the Y {Disease}, making the attacker hard to guess the record owner. But this technique fails when there are records with same disease. Suppose a single patient visit the same hospital with 'r' times and different disease then that patient will have 'r' records but same QID making attacker easy to track. Also, while using k-anonymity and (X, Y)-anonymity, there comes the huge problems of dilemma of choosing the QID. If we take a wrong attribute as QID then the data is either vulnerable or of no use resulting in decrease in data-utility graph.

l-diversity-

Machanavajjhala et al. [10, 14] came with this propose to overcome the shortcoming of attribute linkage which could be easily done in k-anonymity and (X, Y)-anonymity. The main motive of the l-diversity is to prevent attribute linkage so that one is not able to link the sensitive value of one qid group in QID with another sensitive value of another qid group. To maintain a proper distribution of values of sensitive attributes among the qid group as there may be a case one sensitive value may predominate the other values in the sensitive value then attacker can link that value of the table with the value in qid group. So, we maintain an even distribution of the sensitive value using l-diverse value and if the value is well distributed then only it is shown to the user. The notation 'l' in l-diversity requires l diverse values in sensitive values for every value of qid group. This model is also known as p-sensitive k-anonymity model where k=l, as each qid group will contain at least l records. Here, we calculate l-diverse value for every qid group and the least value of the l is considered as l-diversity for the table.

A table is entropy l-diverse, if for each and every qid group is given by following-

$$\sum P(qid, s) \log(P(qid, s)) >= \log(l)$$

Where 'l' is calculated for every qid group in QID and least value is considered as entropy value for the whole table.

Entropy(qid)>=min(entropy(qid_1),entropy(qid_2),……., entropy(qid_n))

But finding the value using this notation is hard to achieve, if the S (sensitive) value occurs too frequently, then it is hard to maintain the ratio between qid and S. 3-anonymous medical dataset as given in Table II, is used to calculate l value

For qid {35-40, professional, male}, where there are 2 patients with cancer and 3 patients with HIV.

-(3/5) (log (3/5) -(2/5) (log (2/5)) =log (1.96)

For qid {30-35, artist, female}, where there are 2 patients with flu and 1 with cancer so entropy for this qid will be

-(2/3) (log2/3) -(1/3) (log (1/3)) =log (1.88)

So according to the rule entropy value for the above table is l<=1.88.

But the problem with the l-diversity is that it prevents attribute record linkage but it does not avoid table linkage.

Also suppose, in a patient table where 95% of them have flu and 5% have HIV. Therefore, for a qid group in QID having record of patient with 50% flu and 50% HIV, the attacker can make a guess with confidence that the record owner is suffering from HIV. However, the above case shows 2-diversity as the percentage of patient having HIV in whole table is 5% but in this qid the percentage is 50% which shows more sign of patient having HIV thereby making him vulnerable.

t-closeness-

As l-diversity, t-closeness works in attribute linkage. t-closeness uses an EMD (Earth Mover Distance) formula to measure the closeness among two distributions of sensitive values in the table [12]. In order to ensure privacy for all patient (record owner) the closeness must be less than t.

But t-closeness has a lot of limitations and weakness which are listed below. As it uses EMD formula, which is not helpful for numerical values, so t-closeness is not a safe approach for the implementing privacy on table with numerical values. Also, t-closeness lacks the ability of providing protection at different levels of sensitive values of the table. Using t-closeness degrades the quality of data resulting in low data-utility value as it asks distribution of data among a qid group that must be same as the distribution of sensitive value among all QID of the table. t-closeness helps us to take the benefits of anonymization other than generalization and suppression of QID, instead of suppressing the whole record one can decide not to show some sensitive records of the row. It depends on choice of individual to hide or show the sensitive attribute of the record, because showing more sensitive records effects privacy and hiding them will degrade the utility of data

δ-Presence –

Using t-closeness, one can only deal with records linkage and attribute linkage. With δ presence, now one can deal with table linkage. Table linkage occurs, if the attacker can confidently refer the presence or absence of the victim record in the released table by the data publisher. The main goal of the privacy provider is to protect data from attacker so that neither he can leak any info from the table nor adversary is able to detect the presence of the victim record from one table to another.

Ercan Nergiz et al. [15] proposed the concept of δ-presence to restrain the probability of referring the presence of any victims record within the range of δ where δ denotes (δ_{min}, δ_{max}).

Suppose one have a general table T1 and private table P2 where T2 is subset of T1 and T1 will satisfy (δmin, δmax) if

δmin <=P(t∈T2|T1) <= δmax, where t∈T1.

δ presence is able to indirectly prevent the record linkage and attribute linkage as if the attacker has δ% surety that victim record in the released table than the chances of linkage of that record to the sensitive record is maximum δ%. δ presence makes an assumption that that attacker can have only access to the table that is owned by the record owner and release by him, with extra knowledge and access. δ-presence may lack to provide enough privacy in the released table. Table III shows the comparison of different anonymity models.

Table III. Comparison of different Anonymity Models

Criteria	k-Anonymity	X-Y Anonymity	*L*-Diversity	t-Closeness	δ-presence
Model Category	Record Linkage	Record Linkage	Attribute Linkage	Attribute Linkage	Table Linkage
Implementation on numerical values	Yes	Yes	Yes	No	Yes
Data Utility (comparison)	Best	Average	Below average	Less	Average
Support for multiple Sensitive attribute	Yes	Yes (but Complex)	Yes	No	Yes
Privacy level (out of 5)	1	2	3	4	5
Risk of Skewness Attack	Yes	Yes	Yes	Yes	No
Requirement of threshold value	No	Yes{k}	Yes/No	Yes	No
Risk of Similarity Attack	Yes	Yes	Yes	No	No

III. CONCLUSION

k- anonymity and (X-Y) anonymity are used to remove record linkage. While l-diversity and t-closeness are used to remove attribute linkage and δ-presence is used to remove the table linkage. Thus, various techniques have been proposed to remove attribute linkage, record linkage, table linkage. The new approach can be thought of where we can combine the previous approaches to come up with the new one having less time and space complexity where the data can be entered in real time and can be searched in minimum of time span. Thus, future efforts could be made in providing the real time searching and insertion of data with less cost of time and space.

REFERENCES

[1] JHS Tomar, JS Kumar, "*A Review on Big Data Mining Methods*", International Journal of Scientific Research in Network Security and Communication, Vol.4, Issue.3, pp.7-14, 2016.

[2] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data mining with big data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[3] H. Taneja, Kapil and A. K. Singh, "Preserving Privacy of Patients based on Re-identification Risk," *4th International Conference on Eco-friendly Computing and Communication Systems (ICECCS)*, vol. 70, pp. 448-454, 2015,.

[4] B. Zhou, J. Pei. And W. Luk,"A brief survey on anonymization technique for privacy preserving publishing of social network data," ACM SIGKDD Explorations Newsletter, ACM New York, NY, USA, vol. 10, no. 2, pp. 12-22, Dec. 2008.

[5] T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," *2008 IEEE 24th International Conference on Data Engineering*, Cancun, pp. 446-455, 2008.

[6] R. Mahesh and T. Meyyappan, "Anonymization technique through record elimination to preserve privacy of published data," *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, Salem, pp. 328-332, 2013.

[7] M. Balusamy and S. Muthusundari, "Data anonymization through generalization using map reduce on cloud," *Proceedings of IEEE International Conference on Computer Communication and Systems ICCCS14*, Chennai, pp. 039-042, 2014.

[8] P. Deivanai, J. J. V. Nayahi and V. Kavitha, "A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data," *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, Tamil Nadu, pp. 732-736, 2011.

[9] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal on uncertainty, fuzziness and knowledge-based systems*, vol. 10 no. 5, pp. 557-570, 2002.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," *22nd International Conference on Data Engineering (ICDE'06)*, Atlanta, GA, USA, pp. 24-24, 2006.

[11] K. Wang and B. C. M. Fung, "Anonymizing Sequential Realease," KDD'06, Philadellphia, Pennsylvania, USA, August 20-23, pp. 1-10, 2006.

[12] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, pp. 106-115, 2007.

[13] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonimity," Proc. 2005 ACM Sigmod Int. Conf. Management of data, Maryland, June14-16, pp. 49-60, 2005.

[14] X. Xiao, K. Yi and Y. Tao, "The Hardness and approxiamation algorithm for l-diversity," Proc. 13th Int. Conf. Extending Database

Technology, Lausanne, Switzerland, March 22-26, pp. 135-146, 2010.

[15] M. E. Nergiz and C. Clifton, "δ-Presence without Complete World Knowledge," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 868-883, June 2010.

Authors Profile

Miss. Gurvinder Pannu working as a Professor; Department of Computer Applications; National Institute of Kurukshetra, India. She has 2 years of teaching experience. She has received her Bachelor of Engineering Degree from Doon valley institute of engineering and technology, Kurukshetra University in 2012. She has received her Master of Engineering Degree from University Institute of engineering & technology, Kurukshetra University in 2014. Her area of Interest is Data Security, Data privacy, Web application security, web attacks, cryptography.

Mr. Shikhar Verma pursued Bachelor of Computer Science from National PG College, Lucknow in 2015. He is currently pursuing Master of Computer Applications from National Institute of Technology, Kurukshetra, India. His main research work focuses on Privacy, Data Security and Algorithms.

Miss. Upasana Arora pursued Bachelor of Computer Science from University of Punjab in 2015. She is currently pursuing Master of Computer Applications from National Institute of Technology, Kurukshetra, India. Her main research work focuses on Big Data and Privacy.

Dr. Ashutosh Kumar Singh working as a Professor and Head; Department of Computer Applications; National Institute of Technology; Kurukshetra, India. He has more than 17 years research and teaching experience in various University systems of the India, UK, Australia and Malaysia. His research area includes Verification, Synthesis, Design and Testing of Digital Circuits. He has published more than 160 research papers till now in peer reviewed journals, conferences and news magazines and in these areas. He is the co-author of six books which includes “Web Spam Detection Application using Neural Network”, “Digital Systems Fundamentals” and “Computer System Organization & Architecture”. He has worked as principal investigator for four sponsored research projects and was a key member on a project from EPSRC (UK) “Logic Verification and Synthesis in New Framework”. Dr. Singh has delivered the invited talks and presented research papers in several countries including Australia, UK, South Korea, China, Thailand, Indonesia, India and USA. He had been entitled for the awards such as Merit Award-03 (Institute of Engineers), Best Poster Presenter-99 in 86th Indian Science Congress held in Chennai, INDIA, Best Paper Presenter of NSC’99 INDIA and Bintulu Development Authority Best Postgraduate Research Paper Award for 2010, 2011, 2012.

