A Comparative Review of Privacy Preservation Techniques in Data Publishing

Atul Kumar
Computer Science Department
MANIT
Bhopal, India
atulatul23.ak@gmail.com

Dr. Manasi Gyanchandani
Computer Science Department
MANIT
Bhopal, India
manasigyanchandani@yahoo.co.in

Priyank Jain
Computer Science Department
MANIT
Bhopal, India
priyankjain88@gmail.com

***Abstract*—Most enterprises generate a huge amount of public and private dataset actively with the integration of modern technology. So,security is a big concern of these dataset. Initially the security is provided at enterprise level but now-a-days it is an inevitable task to provide security at personal level. So to achieve the security generalization, suppression,slicing and one attribute per column slicing is used till now. The aim of this paper is to draw a review of all the existing techniques which are used in privacy preservation with comparative analysis of all anonymization techniques and show the flaw of privacy preservation techniques with respect to different parameters.**

Keywords—Query recommendation, Query Logs, Information Retrieval.

I. Introduction

Agencies and many other organizations frequently need to publish microdata, e.g., health data or survey data, for research purpose and other goals. Typically, such data is collected in a table, and each record (row) belongs to one individual. Each record has certain number of attributes, which can be divided into four categories: -

i. There are some attributes which uniquely identify an individual, by Name or Social Security Number are known as identifiers [1][2].
ii. Attributes that clearly identify individuals. These are called as explicit identifiers which includes Social Security Number, Address, and Name, and so on [1][2].
iii. Attributes whose values when combined together can potentially identify an individual. These are called quasi identifiers and they include attributes like Pin code, Gender and Date of Birth [1][2].
iv. Attributes which do not allow to reveal the data or individual's sensitive information which is not known to the attackers are defined as sensitive attribute, such as Disease and Salary [1][2].

When releasing microdata, it is needed to prevent the sensitive information of the persons from being revealed. The well-known kinds of information disclosure are: -

a) identity disclosure
b) attribute disclosure
c) membership disclosure

Identity disclosure happens when an individual is connected to a specific record in the released table [1]. Attribute disclosure emerges when new data about any individual is revealed, i.e., the distributed information makes it conceivable to make sense of the attributes of an individual more decisively than it would be possible before the information distribute [1]. Identity disclosure leads to re-recognizable proof of an individual and furthermore uncovers the related sensitive values. A spectator of a published table may inaccurately identify an individuals sensitive attribute. She/he then takes a particular value, and behaves according to the perception (become aware of something through the senses). This can harm the individuals data, even if the perception is wrong. The attacker can get the participation data of any person by checking the nearness of QI value in bucketized data. If the QI value is absent in the bucketized information, it implies that the individual isn't in the original data.

Membership disclosure occurs when the data set that has to be published is chosen from a huge population and the selection criteria are sensitive attributes (e.g., only diabetes patients are selected). The published table which gives useful information to attackers and presents the risk of disclosure or we can say that it discloses the information about individual.

So, the main objective in data publishing is to limit the disclosure risk to a threshold level by optimizing the benefit as well as minimizing the information loss. This is accomplished by anonymizing (generalizing, suppressing, slicing) the data prior to publishing. A familiar anonymization approach is generalization, which substitutes QI values with values that are semantically consistent but less specific. As a result, more records will have the same set of QI values. An equivalence class of an anonymized table has been defined which consists a set of records that have the matching attributes for the QI. The flow of data publishing is shown in figure fig:1

978-1-5386-0807-4/18/$31.00 ©2018 IEEE

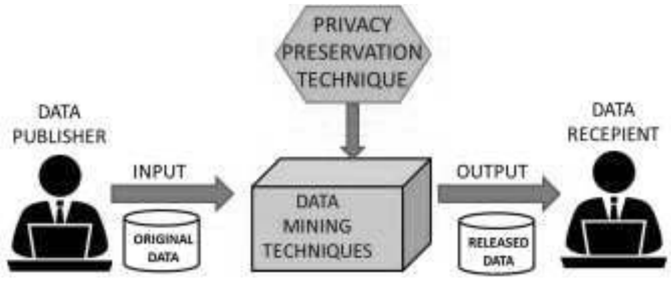

Fig 1: Flow diagram for Data Publishing

Partitioning of database: Data can be partitioned in three different ways as explained below:

a) Horizontal Partitioning: -The data can be subdivided into fragments horizontally where each fragment consists of a subset of the records of relation R.Horizontal partitioning divides a table into several tables with equal number of column and but less number of rows.
b) Vertical Partitioning: -The data can be divided into a set of small physical files each having the subset of the original relation with same number of rows but less number of columns.
c) Mixed Partitioning: -The data is first subdivided horizontally and each subdivided fragment is further partitioned into vertical fragments and vice versa [1][3].

II. Literature survey

To achieve k- anonymity and l diversity in social networks, B.K.Tripathy et. al. used an algorithm in the paper [4]. The algorithm is based upon already existing algorithms developed in this direction. Algorithm is enhanced suitably from their corresponding algorithm for micro data and also depends upon some modified algorithms developed for anonymization against neighborhood attack. The algorithm still needs some enhancements in order to reduce the complexity so that it can be efficiently used on large social networks.

The author of the paper [5] proposed a semantic anonymization approach. The main method is based on the area and the data owner rules to avoid the similarity attacks. The approach modifies privacy preserving techniques to prevent similarity attack. The results show that the semantic anonymization increases the privacy level and suppresses the data utility. . The area based semantic rules and data owner semantic standards are utilized for anonymization process that overcomes the similarity attacks on privacy. The semantic anonymization algorithm needs to be optimized to decrease the information loss and a dynamic version is provided based with a deterministic relation between the utility and the privacy level.

Author presents a (k, l, θ)-diversity model in the paper [6] based on clustering to minimize the information loss as well as assure data quality. Cluster size, distinct sensitive attribute values and the privacy preserving degree has been considered for this model. Extensive experimental evaluation shows that (k,l, θ)-diversity model for privacy preservation dominates the existing approaches in terms of execution time and data utility. The proposed algorithm in paper [6], (k,l, θ)-diversity models against sensitive attribute disclosures. Author hypothetically analyzed the hardness of this problem, and developed efficient algorithms to deal with them. The extensive experiments illustrate that the proposed methods are effective and practical in real world applications. During experiments, it is found that the (k,l, θ)-diversity clustering algorithms orderly dominated the other privacy preservation techniques.
The author evaluates the source of attribute disclosure and suggests a novel idea for privacy protection based on l-Diversity in the paper [7]. Semantic meaning of the sensitive attribute has been taken into consideration and given a stronger definition of privacy protection. First, the sensitive attribute values are divided into groups, and then the records are assembled according to the sensitive attribute. At last, the table is anonymized. While privacy protection becomes more and more important, various solutions have been proposed. K-Anonymity well protects privacy against identity disclosure; however, it doesn’t protect it against attribute disclosure. L-Diversity and (a,k)-Anonymity are proposed to protect privacy from attribute disclosure. T-closeness goes deeper for the distribution of the sensitive attribute values to be close to the distribution in the whole table. But, sufficient protection against disclosure is not provided by these techniques. Author proposed a principle called (a, d)-Diversity based on l-Diversity and takes the real meaning of sensitive attribute and relations among the sensitive attribute values into consideration, since background knowledge is always based on the relationship among the attributes. The advantage of the principle is that it concerns more about the real meaning of the sensitive attributes which the attackers concern the most.

An organized review suggests that numerous methods of anonymization such as generalization and bucketization have been presented for privacy preserving micro data publishing. Latest research has shown that generalization leads to loss of significant information, mainly for high dimensional data. In contrast, bucketization does not avoid membership disclosure. Whereas, slicing offers better data utility than generalization as well as prevents membership disclosure. Slicing overcomes the weaknesses of generalization and bucketization and maintains better data utility while guarding against privacy threats. In slicing, each attribute is in only one column. Overlapped slicing, is an extension of slicing, which replicates an attribute in more than one column [2][8].

III. Existing Methodology And Mechanism

In this section, several mechanisms of privacy preservation are discussed below:

K-anonymization approach: In this approach the problem statement is Given person-specific field-structured data, provides a data release which scientifically assures that the individuals who are the foci of the data cannot be re-identified while the data remain reasonably useful. So to solve this problem K-anonymity came into existence and it can be defined as ”A table fulfills k-anonymity condition if every record in the table is indifferentiable from at least k - 1 other records with respect to every set of QI attributes; such a table is called a k- anonymous table” [4]. Hence, for every combination of the QI values in the k-anonymous table, there are at least k records that have those values.

L-diversity approach: The k -anonymity approach has the lack of diversity so this method(l-diversity) came into existence and it can be defined as ”An equivalence class (EC) satisfies L-diversity if there are at least L well represented values for the sensitive attribute” [5]. L-diversity ensures privacy to sensitive attribute value of a particular person unless the adversary has enough background knowledge to eliminate L-1 sensitive attribute values in the person’s equivalence class. The main benefits of l- diversity are that it restrains homogeneity and background knowledge attack. For example, Data Mining Applications which appears with l-diversity needs every QID (Quasi identifier) group that contains 140 records in Engineering and Medicine with one well represented sensitive values. The knowledge known to attacker and publisher is not same because of which, when knowledge is collected by attacker it is not known to publisher this is the time when l-Diversity comes into picture.

The T-closeness approach: The l-diversity approach has some shortcomings which are, l-diversity is insufficient to prevent attribute disclosure to overcome this problem of insufficiency. T-closeness came into existence and it can be defined as ”An equivalence class is said to have t-closeness if the closeness between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table has maximum threshold *t* ” [6].

There are some more techniques used in privacy preserving are listed as follows [1][7]:-

1. Generalization
2. Multi Set-Based Generalization
3. One-Attribute-Per-Column Slicing
4. Slicing
5. Slicing With Suppression

Let us consider an example of medical record of individual in a hospital to understand these techniques. The original data is shown in table (I).

TABLE I ORIGINAL TABLE

Age	Gender	Zipcode	Disease
22	F	47908	Dengue
22	M	47907	Flu
33	F	47908	Flu
52	F	47907	Migraine
54	M	47303	Flu
60	F	47303	Dengue
60	M	47303	Dengue
64	M	47305	Cancer

1. Generalization:- It is the process of generalizing ranges that substitutes attribute values with more consistent semantically and precise value. It provides the correctness of the data in the record level, but results in less specific information on the k-anonymous dataset. It offers some protection against membership disclosure [1][7]. Table(II) shows the generalized form of original data.

TABLE II GENERALISED TABLE

Age	Gender	Zipcode	Disease
20-52	*	479*	Dengue
20-52	*	479*	Flu
20-52	*	479*	Flu
20-52	*	479*	Migraine
54-64	*	473*	Flu
54-64	*	473*	Dengue
54-64	*	473*	Dengue
54-64	*	473*	Cancer

2. Multi Set-Based Generalization:-In multiset based generalization each column consists of one attribute. The attribute will be in the form [tuple1:frequency,tuple2:frequency.......tupleN:frequncy]. In this approach each attribute preserve exact values because of which association within one bucket can be breached [1][7]. The multi set based generalized data is shown in table (III).

TABLE III MULTI SET BASED GENERALIZED TABLE

Age	Gender	Zipcode	Disease
22:2::33:1::52:1	M:1::F:3	47907:2::47908:2	Dengue
22:2::33:1::52:1	M:1::F:3	47907:2::47908:2	Flu
22:2::33:1::52:1	M:1::F:3	47907:2::47908:2	Flu
22:2::33:1::52:1	M:1::F:3	47907:2::47908:2	Migraine
54:1::60:2::64:1	M:3::F:1	47303:3::47305:1	Flu
54:1::60:2::64:1	M:3::F:1	47303:3::47305:1	Dengue
54:1::60:2::64:1	M:3::F:1	47303:3::47305:1	Dengue
54:1::60:2::64:1	M:3::F:1	47303:3::47305:1	Cancer

3. One-Attribute-Per-Column Slicing:- One attribute per column slicing consists of two terms(one attribute per column + slicing). One attribute per column slicing protects attribute distributional information but each attribute is in its own column therefore it does not preserve attribute correlation. In slicing, one group correlated attributes together in one column and preserves their correlation [1][7]. The one attribute per column sliced table is shown in table(IV).

TABLE IV ONE ATTRIBUTE PER COLUMN SLICED TABLE

Age	**Gender**	**Zipcode**	**Disease**
22	F	47908	Dengue
22	M	47907	Flu
33	F	47908	Flu
52	F	47907	Migraine
54	M	47303	Flu
60	F	47303	Dengue
60	M	47303	Dengue
64	M	47305	Cancer

4. Slicing:-Slicing is an approach which divides the data both horizontally as well as vertically. In horizontal partition tuples gathered into buckets. Inside , section attribute are arbitrarily permuted to part the connecting between various segments. In vertical partition attributes are gathering into sections based on the relationships between the attributes. Every segment contains a subset of highly corresponded attributes. The idea of Slicing is to break the connection between cross segments, however to keep up the connection inside every section. Slicing decreases the dimensionality of the information and preserves great utility than generalization [1][7]. The sliced table is shown in table(V).

TABLE V SLICED TABLE

Age,Gender	**Zipcode,Disease**
22,F 22,M 33,F 52,F	47908,Dengue 47907,Flu 47908,Flu 47907,Bronchitis
54,M 60,F 60,M 64,M	47303,Flu 47303,Dengue 47303,Dengue 47305,Cancer

5. Slicing with suppression: In this technique, partitioning of the dataset is performed both vertically and horizontally. By grouping attributes into columns, vertical partitioning is done based on the attributes' co-relation. As we know that highly correlated attribute includes subset of each column so, horizontal partitioning will group all tuples into buckets. So, to identify the similar QI attribute values in the different tuples and use the suppression technique in the QI field on the tuple. To maintain the privacy in each column interdependence is maintained, but interdependence across columns is not as maintained in suppression slicing. After performing the above method we can preserve privacy results better than generalizations and bucketization. By this approach privacy of identity is protected [1][7]. The slided with suppression table is shown in table (VI).

TABLE VI SLICED WITH SUPPRESSION

Age,Gender	**Zipcode,Disease**
22,F 22,M 33,F 52,F	47908,Dengue 47907,Flu 47908,Flu 47907,Bronchitis
54,M 60,F 60,M 64,M	47303,Flu 473*,Dengue 473*,Dengue 47305,Cancer

An architecture in fig:3 is drawn on the basis of existing methodology.

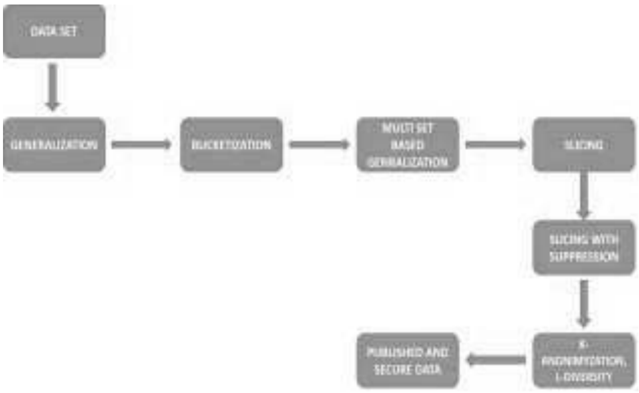
DATA SET
GENERALIZATION
BUCKETIZATION
MULTI SET BASED GENERALIZATION
SLICING
SLICING WITH SUPPRESSION
K-ANONYMIZATION, L-DIVERSITY
PUBLISHED AND SECURE DATA

Fig 2: Architecture based on existing methodologies

IV. COMPARISON OF ANONYMIZATION TECHNIQUES

In this particular comparison TABLE VII, comparison of the anonymization technique on the basis of five parameters is shown:

TABLE VII COMPARATIVE ANALYSIS

Anonymization Techniques	Parameters				
	Reveale d correlati on quantity	Informa tion Loss	Data Type	Priva cy Preser vation	Membersh ip Disclosure
Generalization	High	Very High	Micro Data	Low	High
Multi-Set based generalization	Moderat e	Moderat e	Micro Data	Mode rate	High
One attribute per column slicing	Low	Low	Micro Data	Mode rate	Moderate
Slicing	Very Low	Low	High Dimensiona l	High	Low
Slicing with suppression	Very Low	Low	High Dimensiona l	Very High	Low

Let us try to understand all these five parameters on which comparison has been made:

a) Revealed Co-relation Quality (linkage property):-It can be defined as the ”when two different information of an individual(which are present in table) can be linked together in order to get any personel information”. In samrati example name in a public voter list and once record in a published medical database so by linking these two information malicious attacker can get information what he wants. for example a combination of zip code, date of birth, and sex. From this information Each attribute does not provide uniqueness, but the combination of it can achieve the identity.let us take an another example in which the malicious person noticed that his leader was hospitalized and the record of his leader would be there in the medical patient database so once the attacker reached to the desired patient record so the leader zip code date of birth, and sex, which could serve as the QI in linking attacks.

b) Information loss(data loss):-It is a situation which will be arise when information about a system will get missed by neglecting in storage,transmission and processing failure.

c) Data type(types of data):-

1. Micro data:- Huge amount of data consist of collect home, address, age, employment status, educational level and many other variables which provide separate information for each particular person is called as micro data.
2. High dimensional data:- High-Dimensional data are the data whose extent are enormous than the dimensions (The observation and analysis are considered here are more than one statistical outcome variable at a time). The Example they have considered is Data on health status of patients, which can be high-dimensional (150+ measured/recorded parameters from blood sample analysis, status of the immune system, genetic background, nutrition, alcohol related information).

d) Privacy Preservation:-Privacy preservation is a measure of data security to preserve the identity of the person. There are some privacy preserving approaches in Data perturbation, Data Randomization, Generalization, Manipulation,Sanitation.

e) Membership Disclosure:-we have already discussed it but let it rewind when the breacher will search for the sensitive attribute in the record then in this situation attacker tries to infer membership disclosure.

V. CONCLUSION

Rapid growth in the field of data publishing increases the privacy issues of an individual which expands the capacity of storing and retrieving personal dataset by revealing sensitive attribute information about individuals. Some privacy techniques k-anonymity, l-diversity and T-closeness is well designed for privacy preservation but still they suffer from some privacy attacks. For example K-anonymity is affected by Homogeneity attack and Background knowledge and on the other side L-diversity is affected by skewness attack and Similarity attack. The discussed anonymization techniques Generalization, Multi Set-Based Generalization, One attribute per column slicing, Slicing and Slicing with suppression also have some limitations because slicing and slicing with suppression suffers low membership disclosure. Usefulness of dataset and membership disclosure are important aspects of privacy preservation and anonymization techniques have partially achieved it. so, this survey concludes on that there are a lot of research work is done with the integration of k-anonymity and anonymization techniques but still we can apply these techniques with L-diversity.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy,*Slicing: A New Approach for Privacy Preserving Data Publishing*,IEEE Transaction on knowledge and data engineering vol. 24, no. 3,pp.561-573,March 2012.

[2] Amar Paul Singh, Ms. Dhanshri Parihar, *A Review of Privacy Preserving Data Publishing Technique*,International Journal of Emerging Research in Management & Technology, ISSN: 2278-9359 (Volume-2,Issue-6),pp.32- 36,June 2013.

[3] C. Aggarwal, *On k-Anonymity and the Curse of Dimensionality* Proc. Intl Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.

[4] B. K. Tripathy, Anirban Mitra SCSE, *An Algorithm to achieve kanonymity and l-diversity anonymisation in Social Networks*,Fourth International Conference on Computational Aspects of Social Networks (CASoN),pp.126-130,IEEE 2012.

[5] Emad Elabd, Hatem Abdulkader, Ahmed Mubark,*LDiversity-Based Semantic Anonymaztion for Data Publishing*. Information Technology and Computer Science, 2015, 10, 1-7 Published Online September 2015 in MECS DOI: 10.5815/ijitcs.2015.10.01.

[6] Gaoming Yang, Jingzhao Li, Shunxiang Zhang, Li Yu school of computer science and engineering anhui university of science and technology Huainan, China,*An Enhanced l-Diversity Privacy Preservation*,2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

[7] Qian Wang ,Xiangling Shi ,College of Computer Science, Chongqing University,*(a, d)-Diversity: Privacy Protection Based on l-Diversity*,World Congress on Software Engineering,IEEE,pp.367-372,2009.

[8] Ashwin Machanavajjhala,Johannes Gehrke,Daniel Kifer,*l-Diversity: Privacy Beyond k-Anonymity*,Proc of the International Conference on Data Engineer, Atlanta, 2007.

[9] J. Brickell and V. Shmatikov,*The Cost of Privacy: Destruction of DataMining Utility in Anonymized Data Publishing* Proc. ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys, vol. 42, no. 4, pp. 153, 2010.

[11] X. Jin, M. Zhang, N. Zhang, and G. Das, *Versatile publishing for privacy preservation*, in Proc. ACM KDD, 2010, pp. 353362.

[12] I. Dinur and K. Nissim, *Revealing Information while Preserving Privacy*, Proc. ACM Symp. Principles of Database Systems (PODS),pp. 202-210, 2003.

[13] A. Inan, M. Kantarcioglu, and E. Bertino, *Using Anonymized Data for Classification*, Proc. IEEE 25th Intl Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[14] M. Terrovitis, N. Mamoulis, and P. Kalnis, *Privacy-Preserving Anonymization of Set-Valued Data*, Proc. Intl Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008.