A Review of Privacy Preservation Technique

Avinash Kumar Singh
NRI Institute of Research and Technology, Bhopal
India

Narayan P. Keer
NRI Institute of Research and Technology, Bhopal
India

Anand Motwani
Nri Intitute of Research and Technology,Bhopal.
India

ABSTRACT
Privacy-preserving is one of the most important challenges in a computer world, because of the huge amount of sensitive information on the internet. The paper contains several privacy preservation techniques for data publishing in the real world. There are several privacy attacks are associate but among of them mainly two attacks are record linkage and attribute linkage. Many scientists have proposed methods to preserve the privacy of data publishing such as K-anonymity, ℓ-diversity, *t*-closeness. K-anonymity can prevent the record linkage but unable to protect attribute linkage. ℓ-diversity technique overcomes the drawback of k-anonymity technique but it fail to protect from membership discloser attack. *T-closeness* technique prevents to attribute discloser attack but it fail in identity disclosure attack. Its computational complexity is large. In this paper we present the novel technique call slicing which to be implemented with various data set through prevent the privacy preservation for data publishing. The goals of this paper is re-analysis a number of privacy preservation of data mining technique clearly and then study the advantages and disadvantages of this technique.

Keywords
Privacy preservation, Data publishing, Data security, Pattern Recognition, Data Mining

1. INTRODUCTION
Data mining sometimes called knowledge discovery data (KDD). It is the process of analysing data from different views and summarizing it into new advance useful information. Today, data mining is used by many organization in different areas like retail marketing, financial, healthcare, communication, and global marketing. Extraction of hidden foretelling information from bulky databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses and big data. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, etc., are used for knowledge discovery from databases. BENJAMIN C. M. FUNG explain in this pater [4] that the data publisher has a record which contain several identifier like Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes, where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners, Quasi Identifier (QID) is a set of attributes that could potentially identify record owners, Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories . All the four identifier exist in each record and every identifier pay important role in privacy preservation. Now we are going to focus on different technique in data mining which to be explain below one by one.

2. K-ANONYMITY
When we discus about the publishing the micro data then we think about several types of risks are associated with the data. To limit the risk of occurrence, Sweeney [14] provides for the confidentiality of k-anonymity, which requires that each record in the table standard does not stand out at least k-1 other records within the data set towards a set of quasi-identifier attributes. There are commonly two methods like generalization and suppression mainly used to protect the privacy preservation in data mining. Bee-Chung Chen in [13] demonstrated that releasing a data table by simply removing identifiers (e.g., names and social security numbers) can seriously breach the privacy of individuals whose data are in the table. There are two attacks in k-anonymity. Homogeneity attack and the background knowledge attack. In homogeneity attack a group of member has same set and has a similar attribute, it may a chance of leak the information. Background knowledge attack it another type of attack in which attacker who has past information about particular record on behalf he can easily guess about the particular records. To protect from those attack we gave to used generalized and suppression methods. The original table given in which several attribute like name, sex, age and zip code. If we publish the table any can who know easily. That's why we used the generalized table to publish data.

Table 1: Original micro data of voter registration list

Name	Gender	Age	Zip code
Ann	Male	30	13568
Bob	Male	32	13526
Carol	Male	35	13566
John	Female	38	13588
Cary	Female	39	13255

In the generalization technique first of all removes explicit identifier and suppressed the quasi identifier age from age 30 to age 39, so that nobody can guess about the quasi and sensitive identifier Information. Second, it is very complex task for the analyst to analysis about the data due to the equal distribution of data. Third, due to each attribute is generalized separately, correlations between different attributes are lost.

Table 2: generalized data of voter registration list

Explicit Identifier	Quasi Identifier		Sensitive Identifier
	Sex	**Age**	**Zip Code**
(Ann)	Male	30-39	13***
(Bob)	Male	32-39	13***
(Carol)	Male	35-39	13***
(John)	Female	38-39	13***
(Carry)	Female	39-39	13***

In above table simply removing explicit identifier (Name) that's why nobody can guess about the Ann address and its identity. The generalization of quasi identifier like Zip code and age also help us to save our record. This way we can publish data in secure manner and also preserve identity of individual. But some limitations of k-anonymity technique are as follows. First, it unable to handle micro data due to lose of

3. L-DIVERSITY

Bucketization technique of l-diversity which to be solve the problem of k-anonymity. In Buktetization all separate bucket are created and separate the quasi identifier and sensitive attribute. L-diversity there are two major attacks are associated. Skewness attack and Similarity attack. Skewness attack when overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure. In such type of attack there is a chance that anyone of a similar class can guess 50% being possibility of positive attribute linkage. Similarity attack when the sensitive attribute values in an equivalence class are distinct but semantically similar, than an adversary can learn necessary information from records. The bucketization [9] has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. As shown in, 87 percent of the individuals in the United States can be uniquely identified using only three attributes (Birth date, Sex, and Zip code). A micro data (e.g., census data) usually contains many other attributes besides those three attributes. This means that the membership information of most individuals can be inferred from the bucketized table.

Table 3: Bucketized data of voter registration list

	Sex	**Age**	**Zip code**	**BID**
(Ann)	Male	30	13053	1
(Bob)	Male	32	13053	1
(joy)	Male	35	13053	1
(Joni)	Female	38	13054	2
(Ed)	Female	39	13054	2
(joy)	Male	40	13054	2
(John)	Male	50	13055	3
(Rani)	Female	56	13055	3
(Salu)	Male	59	13055	3

Three buckets are created and identified by their bucket IDs (BID). In bucketization technique also attributes are partitioned into columns, one column contains QI values and the other column contains SA values. In bucketization, one separates the QI and SA values by randomly permuting the SA values in each bucket. In some cases we cannot determine the difference between them two. So it has one drawback for micro data publishing. Some limitation of l-diversity is as follows, First, It does not provide membership disclosure attack. When the attributes are less then it success but the attribute are increase like we analysis the census data membership information of most individuals can be inferred from the bucketized table. Second, bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. Third, when we are going to separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

4. T-CLOSENESS

Sometimes, adversary has an information about statistical information that 80% of males in particular university have age above of 42 have cancer. T-Closeness [6] attempts to guarantee privacy against such adversaries by assuming that the distribution of the sensitive attribute (say disease) in the whole table is public information. Privacy is said to be breached when the distribution of the sensitive attribute in an equivalence class is not close to the distribution of the sensitive attribute in the whole table. Also Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian explain in [6] the principle of t-closeness. An equivalence class is said to have t-closeness if the distance between the distributions of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

Table 4: Example of *t*-closeness of medical aspect.

Non Sensitive		**Sensitive**	
Age	**Gender**	**Disease**	**Count**
<42	Male	Flu	400
<42	Male	Flu	200
>42	Male	Cancer	400
>42	Female	Cancer	200
>42	Female	Cancer	400

The t-closeness is a guard against adversary who know marginal distribution of sensitive. Attribute in given record. For example in the above table age less than 42 is disease flu and the age greater than 42 is disease cancer. Some limitation of t-closeness is also exit. First, it lack the flexibility of specifying different protection level for different sensitive values. Second, the EMD function is not suitable for preventing attribute linkage on numerical sensitive attributes. Third, enforcing t-closeness would greatly degrade the data utility because it requires the distribution of sensitive values to be the same in all qid groups. Also explain in paper [4] t-closeness is not suitable method for privacy preservation of publish data. This would significantly damage the correlation between QID and sensitive attributes. One way to decrease the damage is to relax the requirement by adjusting the thresholds with the increased risk of skewness attack.

5. RANDOMIZATION

Randomization technique is used to hide the data by adding the noise and adding the mask with the table. The noise added is rightfully large so that the individual values of the records can no longer be recovered by any adversary. Also explain by Pingshui Wang in [16] In general, randomization method aims at finding an appropriate balance between privacy preservation and knowledge discovery. Representative randomization methods include random-noise based perturbation and Randomized Response scheme.

6. SLICING

The new privacy preservation technique which is mostly used, provide more security in compare to the above methods. Slicing, partitions the data both horizontally and vertically. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. It is more secure method in compare to generalization and buketization, because all the following steps are involved in slicing process. Slicing process is operating in following functional steps are given below.

Functional steps:-

Step 1: Extract the data set from the database.

Step 2: Anonymity process divides the records into two.

Step 3: Interchange the sensitive values.

Step 4: Multi set values generated and displayed.

Step5: Attributes are correlate and secure data Displayed.

In a below figure flow of data is going on in secure manner. Also our slicing algorithms are performing in following three steps. Attribute Partition, Tuple partition, and column partition. In the Attribute Partitioning attributes are highly correlated to each other in same column it is nice for utility and privacy. Utility means grouping highly correlated attribute and privacy means associating of uncorrelated attribute which present high degree of identification risk and value of attribute is less frequent. Column Generalization It means how to protect the membership disclosure protection. Actually this phase is not more important that's why it is not required phase, because column generalization is required for membership disclosure protection. If the column value is unique then it has only one matching bucket. The main problem is that unique value is easily identified. In the tuple

partition the main part is that it checks to satisfy the l-diversity or not. In the tuple partition phase used to data structure like queue of bucket Q and sliced bucket Sb.

Table 5: Sliced data

(Age, Sex)	(Zip Code, Disease)
(22, M) (22, F) (33, M) (54, M)	(14589, M) (14586, F) (14587, M) (14588, M)
(56, M) (60, F) (60, M) (64, M)	(145587, M) (14577, F) (14788, M) (14566, M)

In the above table is truly sliced, in which four row in upper row and four row in downward. The attribute of the table are combine in such a way like the attribute (Age, Sex) is correlate with (Zip code, Disease) in 16 possible ways. For example (22, M) can correlate with (14589, M), (14586, F), (14587, M), and row (14588, M). So that we see here the partition of row and column are most secured from the outer world, because nobody can guess what exactly row has a information of personal person.

Slicing architecture is explain in below in which contain several phases from which data flow through step by step. Original table contain original table where data in understandable manner. In second stage data is going to generalized due to secure manner. But due to some problem that table bucketised and its comes in next stage and so that the data comes through various phases it's becomes secured.

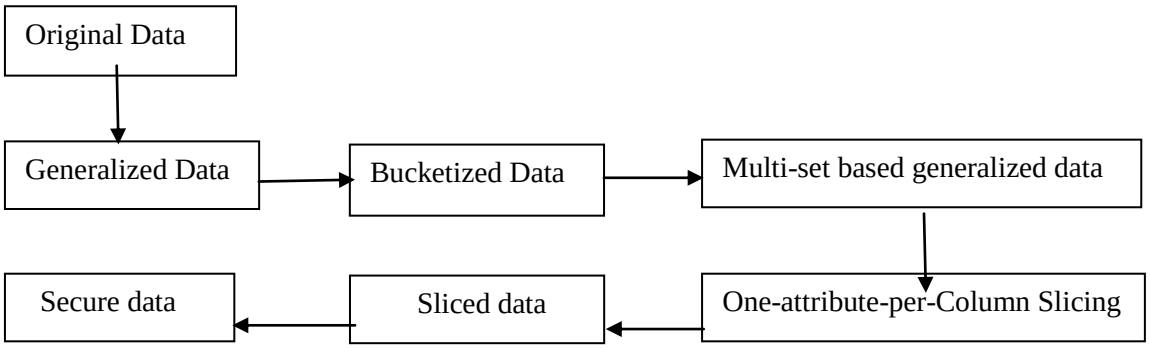

Figure 1: Slicing Architecture

7. DISCUSSIONS AND FUTURE WORK

In this paper, we present a comperative approach of anonymization for providing privacy preservation of data base. Several anonymization methods that is k-anonimity, l-diversity, t-closeness, recdamization, and data slicing. Data Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. We illustrate that how slicing is used to prevent attribute disclosures. The general methodology of this work is before data anonymization one can analyze the data characteristics in data anonymization. The basic idea is one can easily design better anonymization techniques when we know the data perfectly. Finally, we have showed some advantages of data slicing comparing with generalization and bucketization. Data slicing is a promising technique for handling high dimensional data and maintain the high correlation of data and applying permutation and combination over the data for privacy preservation aspect.

8. REFERENCES

[1] R. Mahesh, T. Meyyappan, "Anonymization Technique through Record Elimination to Preserve Privacy of Published Data", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22.Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.

[3] Neha v. Moghre, Sulbha patil, "Slicing: An approach for privacy preservation in high dimensional data using anonymization technique" Proceedings of Fifth IRAJ International Conference, 15th September 2013, Pune, India, ISBN: 978-93-82702-29-0.

[4] Benjamin C. Fung, K E wang, Rui Chen, Philip S Yu "Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.

[5] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam, "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity", International Conference on Data Engineering, 2007, pp106-115.

[6] A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity", In Proc. 22nd Intel international Conference on data engineering. (ICDE), 2006, pp24.Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[7] M.Alphonsa1, V.Anandam2, D.Baswaraj3" Methodology of Privacy Preserving Data Publishing by Data Slicing" International Journal of Computer Science and Mobile Applications, Vol.1 Issue. 3, September-2013, pg. 30-34.

[8] D.Mohanapriya, Dr.T.Meyyappan "High Dimensional Data Handling Technique Using Overlapping Slicing Method for Privacy Preservation" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 6, June 2013.

[9] 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) June 2013.

[10] Neha V. Mogre, Prof. Girish Agarwal, Prof. Pragati Patil "Privacy Preserving for High-dimensional Data using Anonymization Technique" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 6, June 2013 ISSN: 2277 128X.

[11] Neha V. Mogre, Prof. Girish Agarwal, Prof. Pragati Patil" A Review On Data Anonymization Technique For Data Publishing" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[12] Bin Zhou, Jian Pei, WoShun Luk "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data" August 20, 2007.

[13] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala "Privacy-Preserving Data Publishing" Vol. 2, Nos. 1–2 (2009) 1–167.

[14] 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.

[15] Neha Jamdar, Vanita Babane "Survey on Privacy-Preservation in Data Mining Using Slicing Strategy" Volume 2 Issue 11, November 2013.

[16] Pingshui Wang, Jiandong Wang, Xinfeng Zhu ,Jian Jiang "Research on Privacy Preserving Data Mining" 2012 International Conference on Biological and Biomedical Sciences Advances in Biomedical Engineering, Vol.9.