Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Secure Techniques of Data Anonymization for Privacy Preservation

Disha Dubli and D.K Yadav
Department of Computer Applications,
National Institute of Technology Jamshedpur, India

Abstract: Now a day's there is an extensive useof internet, the data present on it should be made available in a way that an individual's privacy is not affected. Recently, many organizations are accumulating gigantic amounts of data which are stored in huge databases. Data publisher gather data from data holders, and publicize this data to data recipient for mining, statistical analysis etc.The released data can reveal secret information of an individual. For providing security to the data, manyanonymization techniques have been designed for privacy preserving and micro data publishing. This paper discusses various anonymization techniques such as generalization, bucketization, slicing and also provide a methodology for enhancing security in the slicing technique.Further, a comparative analysis of the proposed method with existing techniques is discussed.

Keywords:data publishing, data security,data anonymization, privacy preservation, generalization, bucketization, slicing

I. INTRODUCTION

In this age of internet, all the organization collect large volume of information from various sources. For Example, the data collected by government agencies and private companies which can be related to an individual or personal browsing histories of millions of people around the world, collected by web search engines,this type of data isknown as micro-data [1].After collecting data next step is to publish data. The publish data is useful for data mining & research. However, publish data usually contains personal information which may be sensitive, leakage of such sensitive information violates the individual privacy. Examples of some popular recent attacks are example of this like discovering identifying the browsing history of an AOL user [2], disease of the Massachusetts governor [3] etc, because of these attacks, privacy preserving has become an important topic of research. Privacy preserving is important to protect from disclosing the identity of an individual.

A. Privacy Preserving Data Publishing

Microdata contains records which give information about a particular person, organization etc. This data needs to be kept secured, for this many microdata anonymization techniques have been proposed. Some important ones are generalization which is for k-anonymity and bucketizationfor 'l-diversity. In these techniques, attributes are categorized in three types of attributes. They are:

Key attributes. Theseattributes which can uniquely identify an individual, such as Name, Voter ID etc.
Quasi Identifiers QIdAttributes which the adversary may already know and if they are taken together, it can potentially identify an individual, e.g., Birthdate, Sex, and Zipcode
*Sensitive Attributes (SAs)*Attributes, which are unknown to the adversary and are considered sensitive, such as Disease and Salary.

Table 1 shows an example. In this name is key attribute, Gender Zipcode and age is quasi- identifier and disease is sensitive attribute

Name	Gender	Zip code	Age	Disease
Reena	Female	444805	45	TB
Shweta	Female	424806	46	Diabetes
Kavita	Female	424806	58	Fever
Neha	Female	444806	65	Cancer

B. Types of Information Disclosure

The data which is published by the data publisher can disclose a lot of information about an individual therefore the data should be made secured. The various types of information disclosures discussed in the literature [4,5,6] are:
*Identity Disclosure*Identity disclosure occurs when an individual is linked to a particular record in the released data.
*Membership Disclosure*When the data to be published is selected from a larger group and the selection criteria are sensitive then it is important to prevent an adversary from learning whether an individual's record is in the data or not.
*Attribute Disclosure*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 [13]

C. Data Anonymization

The publish data have useful information which contains individual risk. Therefore, the main goal is to maximizing the benefit while minimizing individual risk. Anonymization [12] is an approach for preserving privacy of publish data which seeks to secure the sensitive data and identity of record owners. This can be accomplished by anonymization before publicizing the data. Initially in anonymization, the identity attributes i.e. explicit identifiers are removed. In Table 1, since Name can disclose the identity of a patient, the data owner removes Name from Table and releases it as shown in Table 2.

Table 2: A published data when the adversary has no background knowledge

Gender	Zip code	Age	Disease
Female	444805	45	TB
Female	424806	46	Diabetes
Female	424806	58	Fever
Female	444806	65	Cancer

However, removing identity attributes is not enough, as an attacker may have knowledge of the individualsin publish table. Attacker can gather this information from personal knowledge or from public databases such as voter registration list. From table 2 and table 3 one can deduce that Neha has Cancer.

Table 3: A voter registration list

Name	Gender	Zip code	Age
Reena	Female	444805	45
Shweta	Female	424806	46
Kavita	Female	424806	58
Neha	Female	444806	65

Therefore, it is required to apply proper anonymization techniques so that an individual's privacy is not at risk while the data is being publicized.

II. DATA ANONYMIZATION TECHNIQUES

There are number of techniques for anonymizing the data before it is published. The popular methods are suppression, generalization, bucketization, perturbation and slicing.

A. Suppression

It replaces tuple or attribute values with special symbol "**" that is instead of the original value we replace it with some anonymous value throughout the database .Table 4 is a table generated by suppression in which the value of age and the type of disease is replaced with an "**"

Table 4: A published data by suppression

Gender	Zip code	Age	Disease
Female	444805	**	**
Female	424806	**	**
Female	424806	**	**
Female	444806	**	**

It is easy to implement, only some important values are needed to be replaced using the "*" value but in this the quality of the data drastically reduces.

B. Generalization

Samarati and Sweeney proposed to use generalization [2]. It replaces attribute values with semantically unvarying but less particular value. Due to this replacement, many records have same QI values. Generalization replaces exact values with a more general description to hide the details of attributes, making the QIDs less identifying. If the value is numeric, it may be changed to a range of values. For example, age attribute value 45 can be changed to range 40-60. If the value is a categorical value, it may be changed to another categorical value denoting a broader concept of the original categoricalvalue. Table 5 is table generated by generalization in which the gender value i.e. male/female is changed to person and the age is changed to a range 40-60 .

Table 5: A published table by generalization

Name	Gender	Zip code	Age
Reena	Person	444805	40-60
Shweta	Person	424806	40-60
Kavita	Person	424806	40-60
Neha	Person	444806	40-60

It is easy to find a semantically unvarying value which can be used to implement generalization but it fails on high-dimensional data due to the curse of dimensionality and it also causes too much information loss.

C. Bucketization

It is similar to generalization, but it does not modify any QI attribute or sensitive attribute. Instead, after it divides the records into a number of partitions, it assigns a distinctive ID known as GID to each partition, and all tuple's in this partition are said to have the same GID value. Then, two tables are formed, namely quasi attribute (QI) table and the sensitive table. The anonymize data consist of a set of buckets with permuted sensitive attribute values. Note that the grouping formed by bucketization[7] is equivalent to the grouping formed by generalization, except that bucketization data contains all the original tuple values while generalization data contains some generalized tuple's values. In particular, bucketization has been used for anonymizing high-dimensional data. Bucketization has the advantage of allowing users to obtain the original specific values for data analysis. Bucketization data contains all the original tuple values unlike generalization data.It also allows users to obtain the original specific values for data analysis but it does not prohibit membership disclosure. It also needs a clear difference between QIs and SAs. Table 6 (a, b) shows the idea of Bucketization

Table 6: A published data by bucketization

Gender	Zip code	Age	GID
Female	444805	45	1
Female	424806	46	2
Female	424806	58	3
Female	444806	65	4

a) QId table

Disease	GID
TB	1
Diabetes	2
Fever	3
Cancer	4

b) Sensitive table

D. Perturbation

Under perturbation [8], a value can be changed to any arbitrary value. For example, Male can be changed to Female and vice versa. Table 7 shows an example with perturbation

Table 7: A published data by perturbation

Name	Gender	Zip code	Age
Reena	Male	444805	45
Shweta	Male	424806	46
Kavita	Male	424806	58
Neha	Male	444806	65

E. Slicing

To deal with problems which occurs in generalization and bucketization, T. Li et.al [2012] introduced slicing [9] a new technique to preserve privacy of published data**.** Slicing is a novel data anonymization technique which improves the current state of the art. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuple's into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column.

Table 8: A published data by slicing

Gender , Disease	Zip code, Age
Female , Cancer	444806 ,65
Female , TB	444805,45
Female , Diabetes	424806,46
Female , Fever	424806,58

The dimensionality of the data is reduced by slicing. As compare to generalization and bucketization, preserves better utility. Slicing not only groups highly correlated attributes together but also maintains the correlations between attributes. It breaks the association between uncorrelated attribute, which in turn provide more privacy to publish data. Because these attributes are not rare and identification of this is simple task. Slicing provides better privacy protection because any tuple has more than one multiple matching.Slicing can be effectively used for preventing attribute disclosure.Compare to generalization, Slicing preserves better data utility. Slicing can also deal with high dimensional data.

III. RELATED WORK

Many micro data anonymization techniques have been proposed. The most popular ones are generalization for k-anonymity proposed by L. Sweeney in 2002 [2]. Generalization replaces a value with a "less-specific but semantically consistent" value. Bucketization proposed by X.Xio et.alin 2006 [7] for 'l-diversity first partitions tuple's in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. Recently, several approaches have been proposed to anonymize transactional databases also. Terrovitis et.al in 2008 [10] proposed the km-anonymity model which requires that, for any set of m or less items, the published database contains at least k transactions containing this set of items. Y. Xu et al. in 2008 [11] proposed an approach that combines k-anonymity and 'l-diversity but their approach considers a clear separation of the quasi identifiers and the sensitive attribute. Slicing technique proposed by Tiancheng Li et.al in 2012[9]partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data.

IV. PROPOSED TECHNIQUE

Slicing methodology is one of the best methodology for privacy preserving data publishing as it provides attribute disclosure as well as preserves data utility and it can also work on high dimensional data . But this technique can be made more secure by adding a feature of encryption into it. In this system we will partition the data set both vertically and horizontally and then encrypt the sensitive data attributes so that when the data is published the sensitive value attributes are not easily readableby any user .The method of decryption can be told to only some specific users and then they can use the data, for all the other users the data will be in encrypted form only. This method can enhance the security of the data as an adversary cannot easily identify the value of the sensitive attributes. For example we can encrypt the sensitive attribute disease.

Table 9: A published data by proposed technique

Gender , Disease	Zip code, Age
Female , Dbodfs	444806 ,65
Female , UC	444805,45
Female , Ejbcfuft	424806,46
Female , Gfwfs	424806,58

The proposed technique can enhance the security of the published data. It can be applied to large data sets, it can also be applied in the era of big data. In this the risk of identity and membership disclosure is minimized as it will be difficult for adversary to decipher the sensitive attributes.

Table 10: Comparison between Slicing and Proposed Technique

Parameters	Slicing	Proposed Technique
Quality of Data	High	High
Membership Disclosure	Low	Very low
Identity Disclosure	Low	Very Low
Level of Security	High	Very High

V. CONCLUSION

Anonymization of the data is one of the important method to secure the publish data. Popular techniques of data anonymization like suppression, generalization, bucketization perturbationhave been used for preserving privacy of publish data. There are various constraints with these techniques like suppression reduces the quality of data drastically, generalization is inadequate in handling high dimensional data, bucketization needs to have a clear difference between QIs and SAs, perturbation reduces utility of data. The slicing technique which involves partitioning of data both horizontally and vertically is one of the best methods of anonymization. Compared with generalization, slicing effectively utilize the data and unlike bucketization it prevents membership disclosure. Slicing can also work onhigh dimensional data. However, if we use the proposed technique in which we encrypt a number of sensitive attributes using a good cipher technique, then the security of the published data is enhanced which can prove very beneficial. It can limit the details easily. The various sensitive attributes will not be able to be viewed by any adversary which will reduce the chances of violation of any individual's privacy drastically .Therefore, it is an effective method for privacy preserving data publishing**.**

VI. REFERENCES

[1] P.Samarati and L. Sweeney, “Protecting privacy when disclosing information:-anonymity and its enforcement through generalization and suppression,” SRI International, SRI-CSL-98-04, 1998

[2] M. Barbaro and T. Zeller, “A face is exposed for AOL searcher no. 4417749,” New York Times, 2006.

[3] 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.

[4]G. T. Duncan and D. Lambert, “Disclosure-limited data dissemination,” Journal of The American Statistical Association, pp. 10–28, 1986.

[5] D. Lambert, “Measures of disclosure risk and harm,” Journal of Official Statistics, vol. 9, pp. 313–331, 1993.

[6] M. E. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” in Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pp. 665–676, 2007

[7] X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” Proc. Inٹl Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.

[8] H. Kargupta, S. Datta, Q.Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” in Proceedings of the International Conference on Data Mining (ICDM), p. 99, 2003.

[9] Tiancheng Li, Ninghui Li, “Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012

[10] M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacy-Preserving Anonymization of Set-Valued Data,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008.

[11] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, “Anonymizing Transaction Databases for Publication,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.

[12] C.Saravanabhavan, Dr.R.M.S.Parvathi"An Efficient Approaches for Privacy Preserving In Microdata Publishing Using Slicing and Partitioning Technique"International Journal of Engineering Research and Applications (IJERA), August 2013

[13] Kavita Rodiya and Parmeet Gill, "A Review on Anonymization Techniques for privacy preserving data publishing" IJRET: International Journal of Research in Engineering and Technology, November 2015