ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

A Robust Slicing Technique for privacy preserving of medical data store

Rachitha M.V[1], Aparna R[2], Ruma Panda[3], Nandita Yambem[4]
[1]Student of REVA I.T.M, Department of CSE, REVA I.T.M, Bangalore, India[1]
[2] Associative professor, Department of CSE, REVA I.T.M, Bangalore, India[2]
[3] Assistant professor, Department of CSE, Vemana I.T, Bangalore, India[3]
[4] Assistant professor, Department of ISE, Vemana I.T, Bangalore, India[4]

ABSTRACT: Recent works has shown that several anonymization techniques, like generalization and bucketization, have been designed for privacy preserving microdata publishing. Generalization loses considerable amount of information, especially for high-dimensional data. Bucketization does not prevent membership disclosure and is not applicable for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In this paper, we present a new slicing technique, which partitions the data both horizontally and vertically and preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of the slicing technique is that it can handle high-dimensional data. The new slicing technique can be used for attribute disclosure protection and an efficient algorithm is developed for computing the sliced data that obey the ℓ-diversity requirement. Our experiments confirm that the new slicing technique preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute and also demonstrate that it can be used to prevent membership disclosure

KEYWORDS: Privacy preservation, data anonymization, k-anonymity, l-diversity, data security.

I. INTRODUCTION

The technology that convert clear text into a non-human readable form is called data anonymization.In recent years data anonymization technique for privacy-preserving data publishing of micro-data has received a lot of attention. *Micro-data* contains information about an individual entity, such as a person, a household or an organization. Multiple microdata anonymization techniques have been suggested and the most popular anonymization techniques are Generalization [1,2] for k-anonymity [10]and Bucketization. [3, 4, 5].In each record a number of attributes can be categorized as 1) Identifiers that can uniquely identify an individual, such as Name or Social Security Number.2) some attributes may be Sensitive Attributes (SAs) such as disease and salary and 3) some attributes are Quasi-Identifiers (QI) such as zipcode, age, and sex which may be from publicly available database, whose values, when taken together, can potentially identify an individual. Data anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure. Two widely studied data anonymization technique are generalization and bucketization. The main difference between them is that bucketization does not generalize the QI attributes. Generalization transforms the QI-values in each bucket into “less specific but semantically consistent” values so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, the SAs are separated from the QIs by randomly permuting the SA values in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values. Slicing [6] overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. Slicing illustrate how to prevent attribute disclosure and membership disclosure and preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

II. EXISTING METHODS

ANONYMIZATION TECHNIQUES

Data Anonymization is a process of converting the text data into a non-human readable format. Data anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Detailed data (also called as micro-data) contains information about a person, a household or an organization. Most popular anonymization techniques are Generalization and Bucketization. [7]There are number of attributes in each record which can be categorized as 1) Identifiers such as Name or Social Security Number are the attributes that can be uniquely identify the individuals. 2) some attributes may be Sensitive Attributes(SAs) such as disease and salary and 3) some may be Quasi-Identifiers (QI) such as zip code, age, and sex whose values, when taken together, can potentially identify an individual.

In both generalization and bucketization, one first removes identifiers from the data and then partitions tuples into buckets. In bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket. The anonymized data consists of a set of buckets with permuted sensitive attribute values.

A. Generalization

Generalization transforms the QI-values in each bucket into less specific but semantically consistent values so that tuples in the same bucket cannot be distinguished by their QI values. Three types of encoding schemes have been proposed for generalization:

- Global Recording,
- Regional Recording
- Local Recording.

Global recoding has the property that multiple occurrences of the same value are always replaced by the same generalized value. Regional record is also called multi-dimensional recoding (the Mondrian algorithm) which partitions the domain space into non- intersect regions and data points in the same region are represented by the region they are in. Local recoding does not have the above constraints and allows different occurrences of the same value to be generalized differently.

For example, the month of birth can be replaced by the year of birth which occurs in more records, so that the identification of a specific individual is more difficult. Generalization maintains the correctness of the data at the record level but results in less specific information.

B. Bucketization

Bucketization [8,9] is the process of partitioning tuples 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. The anonymized data consists of a set of buckets with randomly picked sensitive attribute values. Bucketization has been used for anonymizing high dimensional data. But their approach assumes a clear separation between QIs and SAs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

Age	Sex	Zipcode	Disease
21	M	56705	Sinus
21	F	56705	Cancer
32	F	56704	Bronchitis
51	F	56704	Sinus
53	M	56301	Gastritis
59	M	56301	Sinus
59	M	56303	Cancer
63	F	56303	Cancer

Table I- Original Table

Age	Sex	Zipcode	Disease
[20-51]	*	5670*	Cancer
[20-51]	*	5670*	Sinus
[20-51]	*	5670*	Sinus
[20-51]	*	5670*	Bronchitis
[53-63]	*	5630*	Sinus
[53-63]	*	5630*	Cancer
[53-63]	*	5630*	Cancer
[53-63]	*	5630*	Gastritis

Table II- Generalized Table

Age	Sex	Zipcode	Disease
21	M	56705	Cancer
21	F	56705	Sinus
32	F	56704	Sinus
51	F	56704	Bronchitis
53	M	56301	Sinus
59	M	56301	Cancer
59	M	56303	Cancer
63	F	56303	Gastritis

Table III – Bucketizable Table

C. Slicing

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 tuples into buckets. Finally, within each bucket, values in each column are randomly permutated to differentiate between columns. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent but they are identifiable. For an example if the data set contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute.

III. EXISTING SLICING ALGORITHM

Here the algorithm calculates the sliced table T that involves of c columns and gratifies the privacy requisite of ℓ- diversity. This algorithm involves of three steps: attribute partitioning column generalization and tuple partitioning. The three phases are

A. Attribute Partitioning:

In this algorithm attributes are divided such that largely related attributes are in the same column. This is better for utility as well as privacy. With respect to data utility, clustering highly related attributes conserves the relations among those attributes. With respect to privacy, the association of not related attributes shows more identification risks than that of the association of high related attributes since the association of unrelated attribute values is very less common and therefore more identifiable. Thus, it is good to split the associations among uncorrelated attributes to guard privacy. In this step, they have calculated the relations among pairs of attributes and then group attributes on the basis of their correlations.

B. Column Generalization:

Records are generalized to satisfy certain minimum frequency required. They have emphasized that column generalization is not a vital step in their algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

C. *Tuple Partitioning:*

In the tuple partitioning steps, records are divided into buckets. They have changed Mondrian algorithm for tuple partition. Not like Mondrian k-anonymity, no other generalization can be related to the records; they have used of the Mondrian for the reason of dividing tuples into buckets.

D. Membership Disclosure Protection:

Here first inspect how a challenger can conclude membership data from container/storage. Since container liberates the QI values in their real form and more individuals can be solely determined using the QI values, the challenger can easily settle the membership of single individual in the real data by inspecting the regularity of the QI values in the binned information. Precisely, if the regularity is 0, the challenger knows for certain that the individual is not in information. If the regularity is higher than 0, the challenger knows with good assurance that the individual is in the information, since this similar records must fit to that unique as nearly no further individual has the identical values of QI.

E. Sliced Data:

The important advantage of slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub-table with a lower dimensionality. Slicing is also different from the approach of publishing multiple independent sub tables in that these subtables are linked by the buckets in slicing

Age,sex	Zipcode disease
(21,M) (21,F) (32,F) (51,F)	(56704,Sinus) (56705, Cancer.) (56704, Bronchitis.) (56705,sinus)
(53,F) (59,M) (59,M) (63,M)	(56704,Sinus) (56705, Cancer.) (56704,Bronchitis.) (56705,sinus)

Table IV -Sliced Table

IV. PROPOSED WORK

In this paper, a robust slicing technique called *r*-slicing for privacy- preserving data publishing of medical data store is presented. There are several advantages of Slicing when compared with generalization and bucketization. Better data utility than generalization is preserved and there is more attribute correlations with the SAs than bucketization. High-dimensional data and data without a clear separation of QIs and SAs can also be handled.

Slicing, effectively prevents attribute disclosure, based on the privacy requirement of *l*-diversity [11].A notion called *l*-diverse *r*-slicing, is introduced which ensures that the attacker cannot learn the sensitive value of any individual at any cost and the privacy is preserved. We can recover the original data even though the attacker modifies the published original table.

We develop an efficient and robust algorithm for computing the sliced table satisfying the *l*-diversity. The algorithm partitions the attributes into columns, then column generalization is applied, and partitions tuples into buckets. Highly correlated attributes are in the same column; this preserves the correlations between such attributes. The associations between uncorrelated attributes are broken; this provides better privacy as the associations between such attributes are less- frequent and potentially identifying.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

We describe the membership disclosure and explain how *r*-slicing prevents membership disclosure. A bucket of size k can potentially match *kc* tuples where *c* is the number of columns. Because only k of the *kc* tuples are actually in the original data, the existence of the other *kc* − k tuples hides the membership information of tuples in the original data.

r-Slicing partitions the dataset both vertically and horizontally and perform minimization and masking of QI's .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 tuples into buckets. Within each bucket, values in each column are randomly permutated. This break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization.

r-Slicing groups highly correlated attributes together, and preserves the correlations between such attributes and protects privacy as it breaks the associations between uncorrelated attributes, that are infrequent and hence identifying. When the dataset contains QIs and one SA, bucketization has to break their correlation; r-slicing, on the other hand, can group and minimizes some QI attributes with the SA, preserving attribute correlations with the sensitive attribute.

r-Slicing has improved data utility than generalization ands slicing . Additional important benefit of *r*-slicing is that it can manage data with greater dimension. An effective algorithm is developed for calculating the *r* - sliced data complying with the ℓ- diversity requisite. *r* - Slicing provides enhanced utility than generalization and slicing and is more efficient than binning in terms of the sensitive attribute. *r*-Slicing can completely stops membership exposure.

A. r-Slicing Algorithms:

A robust and enhanced *r*-slicing algorithm to obtain ℓ-diverse slicing is introduced. For a given a micro data table T and two factors c and ℓ, the algorithm calculates the sliced table that involves of c columns and gratifies the privacy requisite of ℓ- diversity. Our algorithm involves of five steps: attribute partitioning, column generalization and tuple partitioning, multibased generalization, minimizing and masking generalization. First three steps are similar according to the existing slicing technique. Last two steps are

Multibased Generalization

This is a generalized table where each attribute value is replaced with the multiset of values in the bucket. If the bucket size is large for the QIs then sub-buckets will be formed within each bucket and some of the values will be replaced by closest value for reducing the space complexity.

Minimizing and Masking Generalization

Further, the sliced table can be minimized by omitting QIs for reducing the dimensionality of the data and masking the generalized QI with SA for providing maximum privacy and minimum utility.

V. SIMULATION WORKS/RESULTS

We have simulated our system using JSP, javascript and XHTL and MS-Access for data storage. We have used the following modules in our implementation part. The details of each module for this system are as follows:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

Fig. 1:Original Table

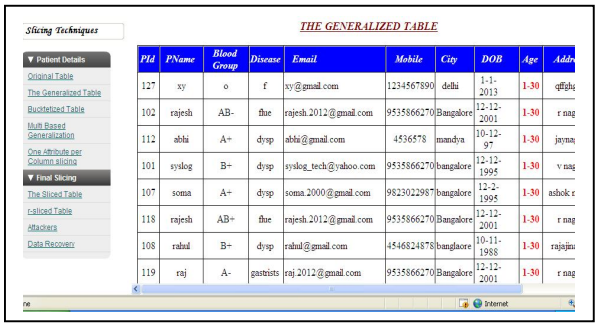

Fig. 2:Generalized TablE

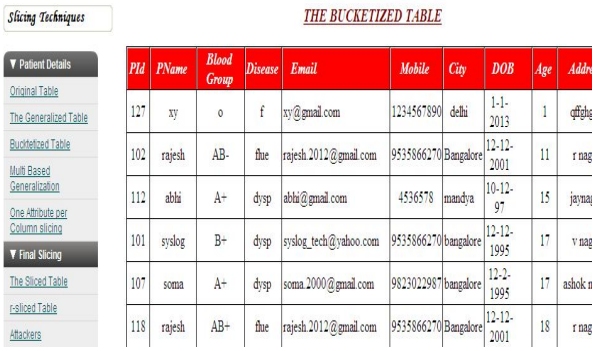

Fig. 3: Bucketized Table

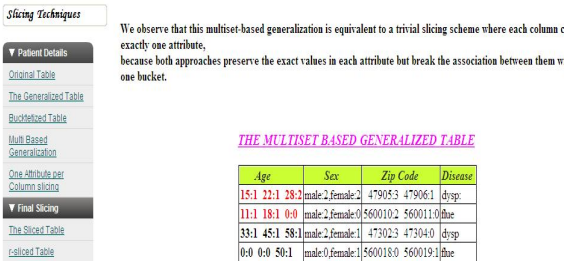

Fig. 4:Multi based Genaralization Table

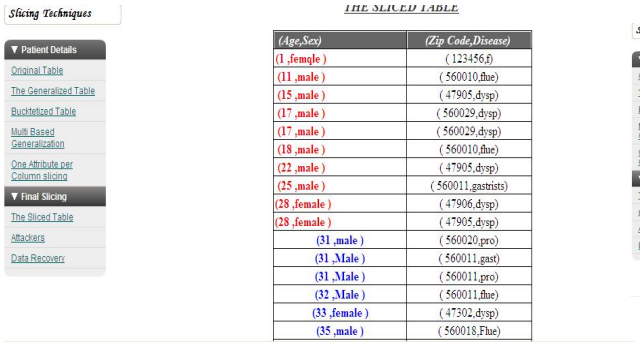

Fig. 5:Sliced Table

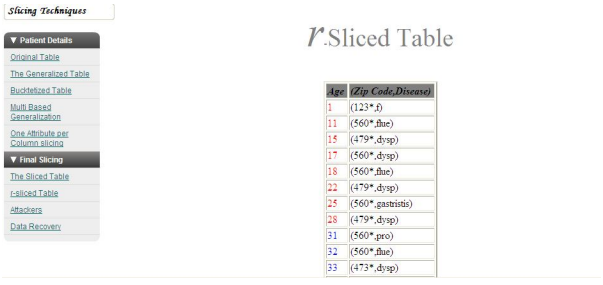

Fig.6 r-Sliced Table

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 2, May 2014

International Conference On Advances in Computer & Communication Engineering (ACCE - 2014)
on 21st & 22nd April 2014, Organized by
Department of CSE & ISE, Vemana Institute of Technology, Bengaluru, India

VI. CONCLUSION

This paper presents a new approach called r-slicing to preserve the privacy of medical data store. *r*-slicing overcomes the limitation of generalization, bucketization,and slicing. Our experiment shows that r-slicing preserves better data utility than the existing algorithms.

REFERENCES

1. P. Samarati. Protecting respondent's privacy in microdata release. TKDE, 13(6):1010–1027, 2001.
2. L. Sweeney. k-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzz., 10(5):557–570, 2002.
3. X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB, pages 139–150, 2006
4. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy- preserving data publishing. In ICDE, pages 126–135, 2007.
5. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang.Aggregate query answering on anonymized tables. In ICDE, pages 116–125, 2007.
6. Tiancheng Li,Ninghui Li,Jian Zhang,Ian Molloy,"Slicing:Anew Approach to Privacy PreseG.ving Data Publishing", March 2012.
7. E. Bertino, D. Lin, W. Jiang (2008). A Survey of Quantification of Privacy. In: Privacy-Preserving Data Mining. Springer US, Vol 34, pp. 183-205.
8. D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy- Preserving Data Publishing," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.*
9. A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity", In Proceedings of PODS'04, pages 223–228, New York,NY, USA, 2004. ACM.K. Elissa, "Title of paper if known," unpublished.
10. L. Sweeney (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol 10(5), pp. 571–588
11.A. Machanavajjhala, J. Gehrke ,D. Kifer, M. Venkitasubramaniam (2007). ℓ-Diversity: Privacy Beyond k- Anonymity. ACM Transactions on Knowledge Discovery from Data, Vol 1(1), Article: 3.

BIOGRAPHY

Rachitha M V, graduated in Bachelor of Technology from VTU, Belguam in computer science & engineering and currently pursing M.Tech specialized in Computer Network Engineering from REVA Institute of Technology .Her current research interest includes Data Mining, Computer networks.

Aparna R, graduated in Master of Technology, and received her M.Tech from VTU. She is associated with REVA ITM as associative professor of CSE department. Her area of interest includes Data ware housing and Data Mining, Design and Analysis **of** Algorithms.

Ruma Panda, graduated in Master of Technology, and received her M.Tech from West Bengal University of Technology and currently pursuing Ph.d from VTU. She is associated with VEMANA IT as Assistant professor of CSE .Her area of interest includes Data Mining, Design and Analysis **of** Algorithms, Formal Languages and Automata Theorem, Compiler Design.

Nandita Yambem, graduated in Master of Technology, and received her M.Tech from Tezpur University ,Assam , and currently pursuing Ph.d from Jain University. She is associated with VEMANA IT as Assistant professor of ISE .Her area of interest includes Data Mining, Design and Analysis **of** Algorithms, Cryptography and Web Services.

.