ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 6, Issue 12, December 2017**

# Effective than Bucketization for the Sensitive Attribute Data Publication

Dr. V.Vijayadeepa, D.Deepika

Associate Professor, Department of Computer Science, Muthayammal College of Arts & science, Rasipuram, Namakkal DT, India

Department of Computer Science, Muthayammal College of Arts & science, Rasipuram, Namakkal DT, India

**ABSTRACT:** Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from disclosure. There are anonymization techniques, like generalization and bucketization, have been designed for privacy preserving microdata publishing. But generalization loses considerable amount of information, for high-dimensional data. Bucketization does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. A new novel technique called slicing, which partitions the data both horizontally and vertically. Overlapping slicing, this duplicates an attribute in more than one column. Overlap slicing preserves better utility than generalization and is more effective than bucketization for the sensitive attribute and used to prevent membership disclosure.

**KEYWORDS:** Data anonymization, overlapping slicing, Data Privacy

## I. INTRODUCTION

Privacy Preserving publishing of microdata has been studied extensively in recent years. Microdata contain records each of which contains information about an individual entity, such as a person, a household, or an organization. Several microdata anonymization techniques have been proposed. For anonymization the attributes are of three categories 1) identifiers used to identify an individual. E.g.: username, 2) Quasi Identifiers are public to an individual and when they are linked to the other databases E.g.: gender, birth date 3) Sensitive attributes which must be protected and kept in privacy E.g.: disease, phone number.

In the data collection phase, the data holder collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data holder releases the collected data to a data miner or the public, called the data recipient, who will then conduct data mining on the published data. Fig-1 denotes Data collection and Data Publishing phase for the data analyzer to from the data owner.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

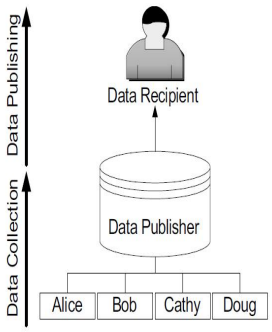


Fig-1: Data collection and Data Publishing

An individual's name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth, and sex, as shown in Fig-2. Each of these attributes does not uniquely identify a record owner, but their combination, called the quasi-identifier, often singles out a unique or a small number of record owners.

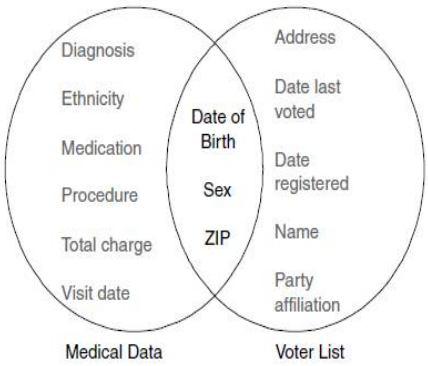


Fig-2: Linking to re-identify record owner

The released data gives useful information to researchers; it presents disclosure risk to the individuals whose data are in the released data. Therefore, the objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the release of data.

## II .EXISTING SYSTEM

In this generalization and bucketization have been designed for privacy preserving micro data publishing.

**Generalization**

In generalization first removes identifiers from the data and then partitions tuples into buckets. It 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.

The original table Table-1 which is to be published is the micro data which is to be anonymized before publishing it. It will contain identifiers, quasi identifiers, and sensitive attributes.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 22 | M | 47906 | Dyspepsia |
| 22 | F | 47906 | Flu |
| 33 | F | 47905 | Flu |
| 52 | F | 47905 | Bronchitis |
| 54 | M | 47302 | Flu |
| 60 | M | 47302 | Dyspepsia |
| 60 | M | 47304 | Dyspepsia |
| 64 | F | 47304 | Gastritis |

Table-1: Microdata to be published

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [20-52] | * | 4790* | dyspepsia |
| [20-52] | * | 4790* | flu |
| [20-52] | * | 4790* | flu |
| [20-52] | * | 4790* | bronchitis |
| [54-64] | * | 4730* | flu |
| [54-64] | * | 4730* | dyspepsia |
| [54-64] | * | 4730* | dyspepsia |
| [54-64] | * | 4730* | gastritis |

Table-2: Generalized Table

The Table-2 is the generalized table from the original table. But generalization loses considerable amount of information, for high-dimensional data.

**Bucketization**

Bucketization does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In bucketization first removes identifiers from the data and then partitions tuples into buckets.

It separates the SAs 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. Bucketization first partitions 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. Table-3 is the bucketized Table from the original table.

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 22 | M | 47906 | flu |
| 22 | F | 47906 | dyspepsia |
| 33 | F | 47905 | bronchitis |
| 52 | F | 47905 | flu |
| 54 | M | 47302 | gastritis |
| 60 | M | 47302 | flu |
| 60 | M | 47304 | dyspepsia |
| 64 | F | 47304 | dyspepsia |

Table-3: Bucketized Table

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 12, December 2017

**Disadvantages of existing system**

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. 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. Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data

## III. PROPOSED SYSTEM

Slicing is technique used for privacy-preserving during data publishing. It is a data anonymization technique which 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 (or sorted) to break the linking between different columns.

The overlapping slicing method for handling high into more than one column is used. It provides better data utility using l diversity requirement for high dimensional data. It use an efficient algorithm called chi matrix for attribute correlation, to protect privacy by breaking the association of uncorrelated attributes and preserve data utility by preserving the association between highly correlated attributes. This technique releases no correlations between attributes thereby, overlapping slicing preserves better data utility in workloads involving the sensitive attributes.

**Advantage of Proposed System**

It preserves better data utility than generalization. It preserves more attribute correlations with the SAs than bucketization. It can also handle high-dimensional data and data without a clear separation of QIs and SAs. Slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of 'l-diversity. Overlapping slicing duplicates an attribute in more than one column. It releases more attribute correlation.

## IV.CONCLUSION

Slicing partitions the data both horizontally and vertically. It preserves better data utility than generalization and preserves attribute correlations with the SAs than bucketization. Data slicing preserves better utility than generalization and also does not requires clear separation between Quasi-identifying and sensitive attributes. The sliced table can be computed efficiently. It can also handle high-dimensional data and data without a clear separation of QIs and SAs. It breaks association between the attributes in that data. Overlapping slicing, this duplicates an attribute in more than one column. These releases more attribute correlations.

## REFERENCES

1. Aggarwal, C. (2005) 'On k-Anonymity and the Curse of Dimensionality', Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909.
2. Ghinita, G. and Kalnis, P. (2008) ' On the Anonymization of Sparse High-Dimensional Data' Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724.
3. Nan Zhang and Wei Zhao (2011) "Privacy-preserving OLAP: An Information-theoretic Approach" Vol. 23, No. 1.
4. Tiancheng Li and Ian Molloy (2012) 'Slicing: A New Approach for Privacy Preserving Data Publishing ' Vol. 24, No.
5. Venkatasubramanian, S. and Li, T. (2007) 't-closeness: Privacy Beyond K-anonymity And 'l–Diversity', Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115.
6. Yaping Li, Minghua Chen, Qiwei Li, And Wei Zhang, (2012) 'Enabling Multilevel Trust In Privacy Preserving' Data Mining Vol. 24, No. 9.
7. Zhang, J. and Li, N. (2009) 'Modeling and Integrating Background Knowledge in Data Anonymization' , Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 6-17.