Identity Disclosure Protection in Slicing for Privacy Preservation

M. Senthil Raja, D. Vidyabharathi

Abstract*—. In recent years privacy preservation micro data publishing has gained wide popularity. Two of the most widely used anonymization techniques are generalization and bucketization. Bucketization doesn't prevent membership disclosure and it doesn't apply for data that don't have a clear distinction between quasi-identifiers and sensitive attribute. On the other hand, generalization loses high amount of data. A combination of both i.e., slicing provides better data utility but still its prone to attacks. Slicing protects the data against membership and attribute disclosure but it doesn't provide any details about identity disclosure. To overcome this we apply k-anonymity through ranging which will improve the overall utility and privacy of data. Here the data is not lost as well as it doesn't result in inference attacks.

***Index Terms*— Anonymization, Data Privacy, Privacy Preservation, Slicing.**

I. INTRODUCTION

Data privacy is one of the most interesting as well as a highly developing research field in data mining. While many of the researchers are concentrated on data mining, privacy is also one of the key concerns. The data which is published by the dataset or any organization should maintain anonymity of the users whose sensitive attributes are shared.

Publishing of data with privacy has been a research issue of wide importance in the recent years. Generally micro data is made up of 3 factors viz., identifier, quasi-identifier and sensitive attribute. There are two basic approaches for privacy preservation viz. generalization and bucketization. In generalization the quasi identifiers are hidden i.e. removed. This leads to the data loss. Bucketization on the other hand preserves data utility but lacks from membership disclosure protection.

The existing system slicing [1] prevents the micro data from membership disclosure and attribute disclosure in a few days cases, however if the attacker suspects about a persons' sensitive attribute and hacks the data publishing then he/she may by pass and breach the identity disclosure protection provided through attribute disclosure protection.

II. RELATED WORK

Slicing [1] partitions the data both vertically and horizontally. Data utility is highly preserved and protection against membership disclosure is also done. One more aspect of slicing is that it handles huge dimensional data. The results of [1] prove that slicing preserves better utility than generalization and are more effective than bucketization. However, one drawbacks of slicing is that it's little unclear about how identity disclosure should be defined for sliced data.

[2] discusses about the dimensionality in k-anonymity. In many cases, users will be interested in disclosing certain sensitive information about them only if the privacy of their data is preserved. [2] analyses k-anonymity approach for the high dimensional case.

Sub linear Queries' (SuLQ) notion of privacy and power of privacy is analyzed in [3]. Sub linear Queries are very important but are generally not concentrated much upon.

[4] discusses about how data utility is reduced is the name of privacy and how there is a tradeoff between privacy and utility. It also discusses how always either of the one are compromised. It surveys on the various methods for improving the utility without affecting privacy.

Privacy Skyline [5] proposes a novel multidimensional approach to quantifying an adversary's external knowledge. This is one of its kinds that discusses about the adversary's knowledge and how to quantify them.

All the above specified works are concentrating on data privacy and utility but doesn't provide solution to inference attacks hence we concentrate on defending inference attacks.

III. MATHEMATICAL PROOF

Step 1: Find the module or operation to which mathematical model is to be proved. The mathematical model is going to be applied to the slicing module which involves l-diversity. Also, we are going to prove measures for correlation.

Step 2: Finding the appropriate measure or model.

a. Slicing:

Once we have computed p(t,b) and p(s|t,b), we are able to compute the probability p(t,s) based on the Bayesian theorem of conditional probability. We can show when t is in the data, the probabilities that t takes a sensitive value sum up to 1.

b. Measures of Correlation:

Two widely used measures of association are Pearson correlation coefficient [6] and mean-square contingency coefficient [6]. Pearson correlation coefficient is used for measuring correlations between two continuous attributes while mean-square contingency coefficient is a chi-square measure of correlation between two categorical attributes. We choose to use the mean-square contingency coefficient because of the fact that most of our attributes are categorical.

Step 3: Create a mathematical formula.

c. Slicing:

For any tuple $t \in D, \sum s$ p(t,s) = 1.

Manuscript Received on July, 2013.

M. Senthil Raja, Computer Science and Engineering, Sona College of Technology Salem, India.
D. Vidhyabharathi, Computer Science and Engineering, Sona College of Technology Salem, India.

$$\sum s\ p(t,s) = \sum s\ \sum B\ p(t,B)\ p(s|t,B)$$
$$= \sum B\ p(t,B) \sum s\ p(s|t,B)$$
$$= \sum B\ p(t,B)$$
$$= 1.$$

l-diverse slicing is based upon the probability p(t,s).

d. Correlation Proof:

Given two attributes A1 and A2 with domains {v11; v12; . . . ; v1d1} and {v21; v22; . . . ; v2d2}, respectively. Their domain sizes are thusd1 and d2, respectively. The mean-square contingency coefficient between A1 and A2 is defined as

$$\varphi^2\ (A1,A2) = \{1/\min(d1,d2)-1\}$$
$$\sum_{i=1}^{d1} \sum_{j=1}^{d2} \{(fij - fi.f.j)^2/fi.f.j\}$$

Here, fi. and f.j are the fraction of occurrences of v1i andv2j in the data respectively. fij is the fraction of co occurrences of v1i and v2j in the data. Therefore, fi and fj are the marginal totals of fij: fi. $=\sum_{j=1}^{d2} fij$ and f.j= $\sum_{i=1}^{d1} fij$. It can be shown that $0 \le \varphi^2 (A1,A2) \le 1$.

For continuous attributes, we first apply discretization to partition the domain of a continuous attribute into intervals and then treat the collection of interval values as a discrete domain. Discretization has been frequently used for decision tree classification, summarization, and frequent item set mining. We use equal-width discretization, which partitions an attribute domain into (some k) equal-sized intervals.

IV. PROPOSED SYSTEM

The proposed system develops a method that involves slicing for data privacy preservation in publishing of datasets for any statistics or survey. In addition we add protection against identity disclosure in the sliced data by applying partial generalization i.e. ranging of the quasi-identifiers.

The following are the list of steps or operations involved in the implementation of our work;

- User interface design
- Data preprocessing
- Anonymity
- Multi set-based generalization
- One-attribute-per-column slicing
- Slicing process

4.1 User Interface Design

The goal of user interface design is to make the user's interaction as simple and efficient as possible, in terms of accomplishing user goals—which often called user-centered design. Good user interface design facilitates finishing the task at hand without drawing unnecessary attention to it. Graphic design may be utilized to support its usability. The design process must balance technical functionality and visual elements (e.g., mental model) to create a system that is not only operational but also usable and adaptable to changing user needs. Interface design is involved in a wide range of projects from computer systems, to cars, to commercial planes; all of these projects involve much of the same basic human interactions yet also require some unique skills and knowledge.

4.2 Data Pre-processing

Data pre-processing is an often neglected but important step in the data mining process. Data gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. In computer science, a preprocessor is a program that processes its input data to produce output that is used as input to another program. The output is said to be a preprocessed form of the input data, which is often used by some subsequent programs like compilers.

4.3 k-anonymity

The *k*-anonymity model is done in order to prevent the re-identification of individuals in the released data set. However, it does not consider the inference relationship from the quasi-identifier to some sensitive attribute. Suppose all tuples in a QID- EC contain the same sensitive value in the released data set, even though the size of the QID- EC is greater than or equal to *k*, all tuples in this QID-EC are linked to this sensitive value in the released data set. Therefore, each individual that has the corresponding QID value will be linked to the sensitive value. Let us call such an attack an inference attack.

There are two possible schemes of generalizations: global recoding and local recoding. With global recoding all values of an attribute come from the same domain level in the hierarchy. In other words, all values come from the values in the same level in the generalization hierarchy.

There are various forms of generalization. In global recoding, a particular attribute value in a domain must be mapped to the same range for all records. In local recoding, different value mappings can be chosen across different anonymized groups. For example if the data is 12 mean we set the data as 10-20, so 12 is hide is hide in 10 to 20. Likewise if data is 1234 mean we set the data as 1000-2000.

4.4 Multi set based generalization

We achieve this by showing that slicing is better than the local recoding approach described as follows. Rather than using a generalized value to replace more specific attribute values, one uses the multiset of exact values in each bucket. The result of using multisets of exact values rather than generalized values improves performance as well as privacy. For the Age attribute of the first bucket, we use the multiset of exact values {22, 22, 33, and 52} rather than the generalized interval [22-52]. The multiset of exact values provides more information about the distribution of values in each attribute than the generalized interval. Therefore, using multi sets of exact values preserves more information than generalization.

4.5 Slicing Process

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 (or sorted) to break the linking between different columns.

A. Figures

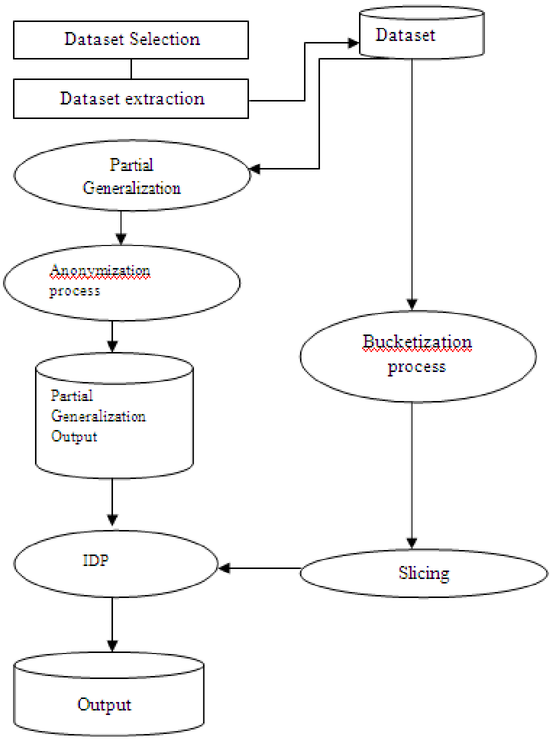

Fig 1: System Architecture

B. Tables

The following depicts the table format of the proposed system.

Table 1: Original Table

(Age, Sex)	(Zip code, Disease)
(23,M)	(636009, Aids)
(36,F)	(636005, Thyroid)
(42,F)	(636009, Malaria)
(42,M)	(636005, Aids)
(57,M)	(636007, Asthma)
(68,F)	(636002, Cancer)
(68,M)	(636002, Flu)
(70,M)	(636007, Flu)

Table 2: Identity Disclosure Protection

Age	Sex	Zip code	Disease
[20-30]	F	[636005-636007]	Aids
[30-40]	M	[636005-636107]	Malaria
[40-50]	M	[636008-636011]	Thyroid
[40-50]	F	[636008-636011]	Aids
[50-60]	M	[636001-636004]	Flu
[60-70]	M	[636001-636004]	Asthma
[60-70]	F	[636006-636009]	Cancer
[65-70]	F	[636006-636009]	Flu

V. CONCLUSION

This work of ours improves data utility as well as preserves privacy. The implementation phase is done through java swing. Java Swing provides an excellent user interface as well as platform independency. This work overcomes three important issues. Firstly it doesn't lose any data instead it provides ranging and swapping of quasi identifiers. Secondly, it provides membership disclosure protection effectively. Finally, our model protects the published data from inference attacks.

The future work on this area can be tried using different datasets. Furthermore instead of normal preprocessing the process of preprocessing can be extended by deploying outlier filtering and over sampling.

ACKNOWLEDGMENT

I thank the Lord Almighty, my parents, professors and friends for their continuous encouragement and support.

REFERENCES

1. Tiancheng Li, Ninghui Li, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, pp. 561-574, MARCH 2012
2. C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.
3. A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SULQ Framework," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.
4. J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.
5. B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.
6. H. Cramt'er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.

AUTHORS PROFILE

M. Senthil Raja is currently pursuing his masters of engineering in the department of Computer Science and Engineering in the Sona College of Technology. He has completed his Bachelors of Engineering also in Computer Science and Engineering. His interests are Data Mining, Data Privacy, and Result Privacy.

D. Vidhyabharathi is Assistant Professor in the department of Computer Science and Engineering in the Sona College of Technology. She has 13+ years of experience in the field of teaching.