ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

A Novel Approach for Secure data Publishing with Membership Disclosure

R.Sravani[1] Kante.Ramesh[2] D.Venkatesh[3]

Research scholar, Dept of CSE, GATES Institute of Technology, Gooty, India[1]

Associate Professor & Head, Dept of CSE, GATES Institute of Technology, Gooty, India[2]

Dean CSE&IT, GATES Institute of Technology, Gooty, India[3]

ABSTRACT: In modern days, for various forms of prepared information include tabular, graph and item set of information, information anonymization techniques have been theme for research. In this paper we present regular review for many forms of several anonymization techniques like generalization and bucketization, have been intended for confidential preserving micro data publishing. Our hot work has presented that generalization lose needed amount of information, particularly for elevated-dimensional data. The hand over, bucketization does not protect member ship disclosure. Where slicing is a technique that preserve better data utility when compare to generalization an also protects member ship disclosure better than bucketization. This paper focus on effective method that can be used s long as better data usage and it can maintain high-dimensional data.

KEYWORDS: Generalization, bucketization, attribute disclosure protection, member ship disclosure protection, slicing

I INTRODUCTION

Confidential-preserving micro data publishing has been studied extensively in modern days. Micro data contain records each of which contains data about and personalize entity, such as a person, hospital, or an company. Various micro data anonymization techniques have been introduced. The most familiar ones are generalization for k-anonymity and bucketization for ℓ-diversity. In two techniques, attributes are of three categories: (1) some attributes are identifiers that can distinctively identify an individual, such as Name or common Security Number; (2) some attributes are Quasi-Identifiers (QI), which the opponent may already know (may be from other publicly-available information) and which, when taken together, can potentially identify an individual, e.g., Birth- date, Sex, name and Zip code; (3) some attributes are Sensitive Attributes (SAs), which are unidentified by an opponent and are measured sensitive, such as type of account, Disease and Salary.

In two techniques such as generalization and bucketization, one initially removes identifiers from the data and then partitions tuples into buckets. The two techniques differ in the next step. Generalization transform the QI-values in each bucket into "less detailed but meaningful context" values so that tuples in the same bucket cannot be distinguished by their QI values. 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:

The generalization procedure depends on attributes or application involved, a user may choose some attributes to remain at a rather low abstraction level while others are generalized to higher levels. The control of how high an attribute should be generalized is normally quite subjective. The control of this procedure is called **"Attribute generalization control".**

There are many possible ways to control a generalization procedure among them we used two common procedures.

The first technique, called**" Attribute generalization threshold control",** either sets one generalization threshold for all of the attributes, or sets one threshold for each attribute. If the number of unique values in an attribute is greater than the attribute threshold, further attribute removal or attribute generalization should be performed. Data mining systems normally allow default attribute threshold value ranging from 2 to 8 and should allow expert and user to modify the threshold values as well. If a user feels that the generalization reaches so high a level for a particular attribute, the threshold can be increased. This corresponds to drilling down along the attribute. Also, to further

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

generalization a relation, the user can reduce the threshold of a particular attribute, which corresponds to rolling up along the attribute.

The second technique, called **"Generalized relation threshold control",** sets a threshold for generalization relation. If the no. of distinct tuples in the generalized relation is greater than the threshold, further generalization should be performed. Otherwise, no further generalization should be performed. Such a threshold should be present in data mining systems usually ranging from 10 to 30 or set by expert or user and should be adjustable. For example, if a user feels that the generalized relation is so small, he can increase the threshold, which implies drilling down. Otherwise, to further generalize relation, he can reduce threshold, which implies rolling up.

These two techniques can be applied in sequence; first apply the attribute threshold control technique to generalize each attribute, and then apply relation threshold to further reduce the size of the generalized relation. No matter which generalization control technique is applied, the user should be allowed to adjust the generalization thresholds in order to obtain interesting concept descriptions.

II RELATED WORK

Slicing is a new approach for secure data publishing with member ship disclosure protection. It can reduce the dimensionality of data in large data base by following the method that it cans partition the data both horizontally and vertically. Horizontal partition is done by grouping the tuples into buckets and vertical partition is done by grouping the attributes in to columns based on correlation among attributes. When compare to generalization and bucketization it can provide better data utility.

Table 1 shows an example micro data table and its anonymized versions using various anonymization techniques. The original table is shown in 1(a). The QI values are {c-age, c-sex, zip code}, and the sensitive attribute SA is type of account. The generalized table for 4-anonymity is shown in table 1(b), a bucketized table for 2-diversity is shown in table 1(c), a generalized table by replacing the multi set based generalization is shown in table 1(d) and two sliced tables are shown in table 1(e) and table 1(f).

Slicing first partitions attributes in to columns each column contains subset of attributes. The horizontal partition is done by grouping the tuples into buckets. The vertical partitions the table. For example, the sliced table in table 1(f) contains 2 columns i.e., {c-age, c-sex} and {zip code and type of account} whereas in table 1(e) contains 4 columns each attribute is exactly at one column.

Slicing also partitions tuples in to buckets. Each bucket contains a subset of tuples. This horizontal partitions the table in table 1(e) contains 2 buckets, each containing 4 tuples.

The main criteria of slicing are it can partitions the tuples in to buckets by randomly permutated to break the linking between different columns.

For example in table 1(f). The values {(24,M),(24,F),(33,F),(44,F) are randomly permutated and there values are {(1234903,SAVING),(12345904,CURRENT),(12345904,LOAN),(12345903,SAVING)} are randomly permutated so that the linking between two columns within one bucket is hidden.

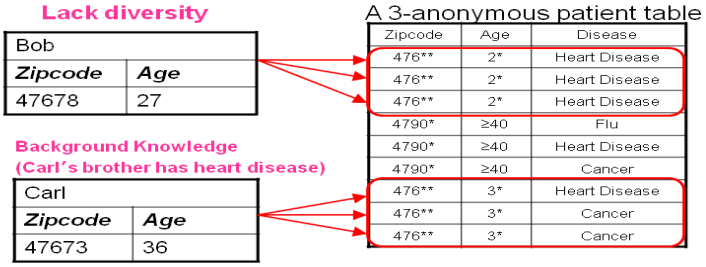
The sliced table is shown in table 1(f) satisfies 2-diversity consider a tuple t1 with QI values as {24, M, 12345903) for determining the t1's sensitive value consider the matching buckets in table 1(f) by observing first bucket B1 t1 must be in B1 because there is no matches of (24, M) in bucket B2 therefore it is conclude that t1 must be in B1.

The next attribute is ZIPCODE attribute it is in second column as (zip code, type of account) in bucket B1 we know that column value for t1 is(12345903,CURRENT) and (12345903,SAVING) are two possible values for t1 sensitive values no other columns have 12345903 as zip code. CURRENT ACCOUNT and SAVING ACCOUNT are possible values for tuple t1.

A. System architecture: The basic data used in this paper is

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013



III PROTECTION AGAINST IDENTITY DISCLOSURE

Protection against identity disclosure guarantees that adversaries will not be able to associate specific records with a known individual. The most popular guaranty is *k*-anonymity [Samarati 2001, Sweeney 2002]. *K-anonymity* guarantees that each record will be indistinguishable from other *k*-1 records, with respect to the quasi identifiers. Every combination of quasi identifiers appears 0 or more than *k* times in the anonymized dataset.

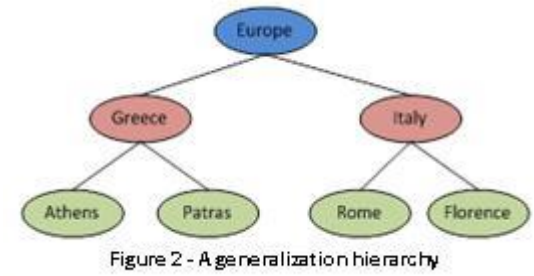


Figure 2 - A generalization hierarchy

In [He et. al. 2009] a method for transforming set-valued data to a k-anonymous form, the *Partition* algorithm, is proposed. *Partition* employs generalization for transforming the data to k-anonymous form. Generalization is the replacement of a group of original values by one new more abstract one. For example, if the residence area of an individual is reported in terms of cities, the city name can be replaced by the country name in anonymized data as in Figure 2. *Partition* employs local recoding; not all appearances of an original value are replaced by a generalized one. *Partition* is a top down algorithm; it starts by considering that all values are generalized to the more generic value of the generalization hierarchy and then drills down the hierarchy until the *k*-anonymity property no longer holds.

CAGE	CSEX	ZIPCODE	TYPE OF ACCOUNT
24	M	12345903	CURRENT
24	F	12345903	SAVING
33	F	12345904	SAVING
44	F	12345904	LOAN
55	M	12345905	SAVING
65	M	12345905	CURRENT
65	M	12345906	CURRENT
64	F	12345906	FIXED DEPOSIT

a) The Original Table

C-AGE	C-SEX	*ZIPCODE*	TYPE OF ACCOUNT
{20-50}	*	1234590*	CURRENT
{20-50}	*	1234590*	SAVING
{20-50}	*	1234590*	SAVING
{20-50}	*	1234590*	LOAN
{55-65}	*	1234590*	SAVING
{55-65}	*	1234590*	CURRENT
{55-65}	*	1234590*	CURRENT
{55-65}	*	1234590*	FIXED DEPOSIT

b) The Generalized table

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

CAGE	CSEX	ZIPCODE	TYPE OF ACCOUNT
24	M	12345903	SAVING
24	F	12345903	CURRENT
33	F	12345904	LOAN
44	F	12345904	SAVING
55	M	12345905	FIXED DEPOSIT
65	M	12345905	SAVING
65	M	12345906	CURRENT
64	F	12345906	LOAN

c) The Bucketized Table

C-AGE	C-SEX	ZIPCODE	TYPE OF ACCOUNT
{24:2 33:1 44:1}	M:1 F:3	{12345903:2 12345904:2}	CURRENT
{24:2 33:1 44:1}	M:1 F:3	{12345903:2 12345904:2}	SAVING
{24:2 33:1 44:1}	M:1 F:3	{12345903:2 12345904:2}	SAVING
{24:2 33:1 44:1}	M:1 F:3	{12345903:2 12345904:2}	LOAN
{55:1 65:2 64:1}	M:3 F:1	{12345905:2 12345906:2}	SAVING
{55:1 65:2 64:1}	M:3 F:1	{12345905:2 12345906:2}	CURRENT
{55:1 65:2 64:1}	M:3 F:1	{12345905:2 12345906:2}	CURRENT
{55:1 65:2 64:1}	M:3 F:1	{12345905:2 12345906:2}	FIXED DEPOSIT

d) Multi set Based Generalization

C-AGE	CSEX	ZIPCODE	TYPE OF ACCOUNT
24	M	12345903	SAVING
24	F	12345903	SAVING
33	F	12345904	CURRENT
44	F	12345904	LOAN
55	M	12345905	CURRENT
65	F	12345905	FIXED DEPOSIT
65	M	12345906	CURRENT
64	M	12345906	SAVING

e) One Attribute per Column slicing

(C-AGE,C-SEX)	(ZIPCODE,TYPE OF ACCOUNT)
(24,M)	(12345903,SAVING)
(24,F)	(12345903,CURRENT)
(33,F)	(12345904,LOAN)
(44,F)	(12345904,SAVING)
(55,M)	(12345905,FIXED DEPOSIT)
(65,M)	(12345905,SAVING)
(65,M)	(12345906,CURRENT)
(64,F)	(12345906,CURRENT)

f) Sliced data

Table 1: An original micro data table and its anonymized versions using various anonymization techniques

IV ATTRIBUTE DISCLOSURE PROTECTION:

The data values in a dataset are not usually equally important as personal information. A common differentiation in privacy related text is between quasi identifiers and sensitive values. Quasi identifiers are regularly known through several sources and they do not threaten the privacy of an individual. Sensitive values on the other hand, are not considered available through other sources and they expose important personal information. If such a distinction holds and it is known by the data publisher, then data must also be protected against the disclosure of sensitive attributes. A common guaranty for protecting against sensitive values is l-diversity. L-diversity guarantees that any adversary cannot connect her background knowledge with less than we well represented sensitive values. Well-represented is usually defined as a probability threshold: an adversary cannot associate her background knowledge with any sensitive value with probability over 1/l.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

	meat	wine	oranges	Straw berry	Pregnancy test	viagra
vassillia	X	X				
Manolis	X	X	X		X	
Eleni			X			X
Maria		X	X			
kostas	X			X		

Table 2: original data sensitive values are depicted with different color on the right

	meat	wine	oranges	Straw berry	Sensitive Values
vassillia	X	X			Pregnancy Test:1 Viagra:1
Manolis	X	X	X		
Eleni			X		
Maria		X	X		
kostas	X			X	

Table 3: Anonymized data

The first anonymization method that provided protection against attribute disclosure in set-valued attributes. The proposal of relies on separating sensitive values from quasi identifiers as depicted in Tables 4 and 5. The idea of separation was first proposed in the relational context in [Xiao et. al 2006], but it was adjusted and extended in [Ghinita et. al. 2008, Ghinita et. al. 2011] for the set-valued context. The basic idea of proposed the anonymization method is to create clusters of similar records (with respect to quasi identifiers) and then publish at each cluster the quasi identifiers and the sensitive values separately. A benefit of this transformation with respect to generalization and suppression is that it does not require creating groups with identical quasi identifiers. This way the information loss is kept low, even for data of very high cardinality and dimensionality.

IV SLICING ALGORITHMS

A. Attribute partitioning:

Attribute partitioning is done by using attributes that are highly correlated attributes are in the same column. This is very useful technique in terms of both usage and privacy. Grouping highly correlated attributes in to same column can reduce the dimensionality of a data it can increase usage is very easy. The association of uncorrelated attributes can provide identification risks than association of highly correlated attributes because the association between uncorrelated attributes can occur less frequently than association of highly correlated attributes. Therefore it is better to break the associations between uncorrelated attributes, in order to protect privacy.

B. Column Generalization:

Columns are generalized to satisfy the frequency of occurrence. Generalization is nothing but deriving the data from low level to higher level. Bucketization also provides same security level of protection as generalization with respect to attribute disclosure.

Generalization is not a mandatory process although it is useful for identify/member ship disclosure protection. If a column value is unique it is easily identified by adversary a tuple with unique column value can only have matching bucket this method is not a good privacy protection where tuple can belong to one equivalence class/bucket. This is the main problem that unique column value can be identified. So it is useful to apply column generalization to ensure column value appears with at least some frequency.

When column generalization is applied to privacy protection then we have to maintain smaller size buckets because generalization loses considerable amount of information. So it is better for utility by maintaining smaller buckets.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

C. Tuple partitioning:

In tuple partitioning tuples are portioned in to buckets for partitioning tuples in to buckets by using tuple partitioning algorithm.

Algorithm tuple-partition (T, ℓ)
1. Q = {T}; SB = Φ.
2. While Q is not empty
3. Remove the first bucket B from Q; Q = Q − {B}.
4. Split B into two buckets B1 and B2, as in Mondrian.
5. If diversity-check (T, Q ∪ {B1, B2} ∪ SB, ℓ)
6. Q = Q ∪ {B1, B2}.
7. Else SB = SB ∪ {B}.
8. Return SB.

The tuple-partitioning algorithm

The above tuple partitioning algorithm maintains two data structures: (1) a queue of buckets as Q (2) a set of sliced buckets as SB initially SB is empty and Q contain only one bucket including all the tuples. In each iteration the algorithm removes one bucket from Q and splits in to two buckets in SB as sliced buckets the bucket can satisfies l-diversity then it puts these two buckets are put at the end of the for further iterations if the Q becomes empty we have computed the sliced table. The set of sliced buckets are stored in SB.

The criteria of tuple partitioning are l-diversity check it is done by using l-diversity check algorithm.

Algorithm diversity-check (T,T_, ℓ)
1. For each tuple t ε T, L[t] = Φ.
2. For each buckets B in T_
3. Record f (v) for each column value v in bucket B.
4. for each tuple t ε T
5. Calculate p (t,B) and find D(t,B).
6. L[t] = L[t] ∪ {hp (t,B),D(t,B)i}.
7. for each tuple t ε T
8. Calculate p (t, s) for each s based on L[t].
9. If p(t, s) ≥ 1/ℓ, return false.
10. Return true.

V MEMBER SHIP DISCLOSURE PROTECTION

Slicing is a method for privacy protection of member ship information. It is essential that, in the anonymized data a tuple in the original data should have same frequency as tuple which is not in original data. Otherwise adversary can determine by examine their frequencies in anonymized data they can determine the difference between original data tuples from tuples not in original data.

Let us examine how an adversary can determine membership information from bucketization. Bucketization can releases the QI values in their original forms. Adversary can compare bucketized data with QI values of a person by observing voter list information. If the frequency of a matching is zero then an adversary can confirm that individual is not in the data. If the frequency of matching is greater than zero then adversary can confirm that the individual is in the data, because this matching tuple belongs to that individual as almost no other individual has the same QI values.

The general methodology used by slicing for protection of member ship disclosure information is let I be the set of tuples in the original data and I' be the set of tuples that are not in original data and also known as fake tuples if a tuple is in I' it matches at least one tuple bucket in sliced data for member ship disclosure protection slicing consider two measures. The first measure is the number of fake tuples. If the number of fake tuples is zero, the member ship information of every tuple is determined. The second measure is number of matching buckets for original and fake tuples. If matching is similar them member ship information is protected because the adversary cannot differentiate original tuples from fake tuples.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering
Vol. 2, Issue 6, June 2013

VI CONCLUSION AND FUTURE WORK

This work motivates various instructions for future research.

This paper presents a new approach called slicing which is a promising technique for privacy preserving and micro data publishing with member ship disclosure protection. slicing can overcome draw backs of several anonymizing techniques such as generalization and bucketization. It can provide better data utility than bucketization. Slicing is used to protect member ship disclosure information by calculating the frequency of QI values in original and duplicate data. It can reduce the dimensionality of data by combining highly correlated attributes in to one column. Slicing is more effective then bucketization for protecting the information which involves sensitive attribute.

The general method we are used in slicing is that: before anonymizing the data, one can analyze the characteristics and use these characteristics in data anonymizing by combining the highly correlated attributes in to single column by reducing the dimensionality of data.

This work motivates several directions of future research. Slicing is a technique where each attribute is exactly in one column. An extension is that over lapping of slicing which duplicates an attribute in more than one column

Second, we plan to member ship disclosure protection by randomly permuting the sensitive attribute values which is not very effective. We plan to design more effective tuple grouping algorithms.

.REFERENCES

1. P. Samarati. Protecting respondent's privacy in micro data 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. X. Xiao and Y. Tao. Output perturbation with query relaxation. In VLDB, pages 857–869, 2008.

Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In KDD, pages 767–775, 2008.

5. J. Li, Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In SIGMOD, pages 473–486, 2008.
6. D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In SIGMOD, pages 217–228, 2006.

BIOGRAPHY

R.Sravani, graduated in Bachelor of Technology (CSE) from JNTU Anatapur in the year 2010, and currently pursing M.Tech specialized in Computer Science from GATES Institute of Technology, Gooty (Affiliated to JNTU Anantapur). Her area of interest includes data ware Housing & Data Mining.

D.Venkatesh, graduated in Master of Computer Applications, and received his M.Tech from Satya bhama University and currently pursuing Ph.d from Rayalaseema University. He is associated with GATES Institute of Technology as Dean of CSE &IT departments since 2010. His area of interest includes Design and Analysis **of** Algorithms, Computer Networks.

K.Ramesh, graduated from JNTU Anantapur in Bachelor of Technology (Computer Science& Engineering) in the year 2002, M.Tech specialized in Computer Science from JNTU Anantapur in the year 2010 and currently working as Associate Professor & Head of the Department of Computer Science in GATES Institute of Technology .His area of interest includes Data Ware Housing & Data Mining, Formal Languages and Automata Theorem.