Implementation of Slicing for Multiple Column Multiple Attributes: Privacy Preserving Data Publishing

Mr. Tushar S. Dhumal
Computer Science and Engineering
SSGBCOET
Bhusawal (MS), India
E-mail: tushar.s.dhumal@gmail.com

Mr. Yogesh S. Patil
Computer Science and Engineering
SSGBCOET
Bhusawal (MS), India
E-mail: yogeshe146@gmail.com

Abstract — Latest work shows that abstraction loses amount of information for high spatial data. There are several anonymization techniques like Abstraction, Containerization for privacy preserving small data publishing. Bucketization does not avoid enrollment acknowledgment and does not give clear separation between aspects. We are presenting a technique called slicing for multiple columns multiple attributes which partitions the data both horizontally and vertically. We also show that slicing conserves better data service than generalization and bucketization and can be used for enrollment acknowledgment conservation. Slicing can be used for aspect acknowledgment conservation and establishing an efficient algorithm for computing the sliced data that obey the l-diversity requirement Our workload confirm that this technique is used to prevent membership disclosure and it also used to increase the data utility and privacy of a sliced dataset by allowing multiple column multiple attributes slicing while maintaining the prevention of membership disclosure.

Keywords *: Data security, Data publishing, Privacy preservation, Data anonymization.*

*****

I. INTRODUCTION

Privacy-preserving publishing of micro-data has been studied extensively in latest years. Micro-data contain records each of which include information about an individual entity, such as a household, a or an organization. Several micro-data anonymization methods have been proposed. Bucketization for ℓ-diversity and generalization for k-anonymity are the most popular techniques. In both techniques, attributes are partitioned into 3 categories: 1. Attributes are identifiers that can uniquely identify an individual, like Age, name or Social Security Number 2. Quasi-Identifiers (QI), attributes which are the set of attributes that can be linked with public available datasets to reveal personal identity. e.g., Birth date, Gender, and Zipcode. 3. Sensitive Attributes (SA), which contains personal privacy information, like Disease, political opinion, crime..

Multiple domains contain multiple sensitive attributes, slicing anonymization techniques proposed to prevent the sensitive information. The basic idea of slicing is to split the link cross columns, but to preserve the link within each column. Slicing is in multiple sensitive attributes preserves good usefulness than generalization and bucketization and reduces the dimensionality of the data. Multiple column multiple attributes slicing increase the utility and Privacy of a sliced dataset with multiple sensitive attributes in different domains.

1.2 Contributions & Organization

In paper, we present a new technique called slicing for multiple column multiple attributes for privacy preserving data publishing. Our contributions include the following.

First, we introduce slicing for multiple columns multiple attributes as a new method for privacy preserving data publishing. Slicing has numerous advantages when compared with generalization and bucketization. It conserves better data utility than generalization. It conserves more attribute correlations with the SAs than bucketization. It can also handle high-dimensional data not including a clear Separation of SAs and QIs.

Second, we show that slicing can be effectively used for preventing attribute disclosure, based on the privacy necessity of *l*-diversity. We establish a notion called *l* diverse slicing, which ensures that the adversary cannot learn the sensitive value of any individual with a possibility greater than 1/*l*.

Third, we develop an capable algorithm for computing the sliced table that satisfies *l*-diversity. Our algorithm partition attributes into multiple columns, applies partitions tuples into buckets and column generalization. Attributes that are very much correlated are in the same column; this preserves the correlations between such attributes. The relations between uncorrelated attributes are broken; this provides better privacy as the associations between such attributes are less frequent and potentially identifying.

The rest of this paper is organized as follows: In Section 2, we formalize the related work and compare it with generalization and bucketization. We implementation module in Section 3 and result & discussion in Section 4 And conclude the paper and discuss future research in Section 5.

II. RELATED WORK

The disadvantage of Generalization is it loses some amount of information, specially for high dimensional data. And Bucketization does not prevent membership revelation and it does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes.

Initially proposed on k-anonymity and curse of dimensionality concept. The author proposed privacy preserving anonymization technique where a record is released only if it indistinguishable from k other entities of data. In paper the authors show that when the data contains a large number of attributes which may be considered quasi identifiers, so it becomes difficult to anonymize the data without an unacceptably high amount of information loss. Also the author faced with a choice of either completely suppressing most of

the data or losing the desired level of anonymity. Finally, the work showed that the curse of high dimensionality applies to the problem of privacy preserving data mining. A. Blum [3] proposed a new framework for practical privacy and they named it as SULQ framework. J. Brickell [4] introduced a new anonymization technique called the cost of privacy. In paper they show that query generalization and suppression of quasi-identifiers offer any benefits over trivial sanitization which simply separates quasi-identifiers from sensitive attributes. This work showed that k-anonymous databases can be useful for privacy preservation, but k-anonymization does not guarantee any privacy.

A multi-dimensional method was proposed by B.C. Chen et.[5], which named as Skyline based method. Privacy is important problem in data publishing. I.Dinur [7] proposed another technique of revealing information while preserving privacy. The authors [6] examine the tradeoff between privacy and usability of statistical databases. D.J. Martin, D. Kifer explained [7] that, anonymized data contain set of buckets which is permuted sensitive attribute values. In particular, bucketization used for anonymizing high dimensional data. D.Kifer and J.Gehrke showed that, Slicing has some connections to marginal publication [9]; they have released correlations among a subset of attributes. Slicing is fairly different than marginal publication in a number of aspects. First, marginal publication can be viewed as a special case of slicing which does not having the horizontal partitioning. So, correlations among attributes in different columns are lost in marginal. T P. Samarati proposed two popular anonymizing techniques, generalization and bucketization. Generalization [10], [11], alternates a value with a semantically constant value. D.J. Martin, D. Kifer explained that ,the Bucketization [13], 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. The anonymized data consist of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high dimensional data. The idea of Overlapping Correlation Clustering was suggested by F. Bonchi et al. and can be occupied to the attribute partitioning phase of the slicing algorithm.

Partitioning of Database:-

Data can be partitioned in three different ways that is, like horizontally partitioned data, vertically partitioned data or mixed partitioned data.

Horizontal partitioning: - The data can be partitioned horizontally where each fragment consists of a subset of the records of relation R. Horizontal partitioning [3] [9] [10] [11] divides a table into several tables. The tables have been partitioned in such a way that query references are done by using least number of tables else excessive UNION queries are used to merge the tables sensibly at query time that can affect the performance.

Vertical partitioning: - The data can be divided into a set of small physical files each having the subset of the original relation, the relation is the database transaction that normally requires the subsets of the attributes.

Mixed partitioning: - The data is first partitioned horizontally and each partitioned fragment is further partitioned into vertical fragments and vice versa.

The idea is to build up a well organized method that enables a secure computation along with minimizing the amount of private data that each party discloses to other. Privacy preserving association rule mining may be used to solve these problems for horizontally partitioned database.

III. IMPLEMENTATION DETAILS

A. *Slicing:*

Slicing divides the data both horizontally as well as vertically. In horizontal partition tuples grouped into buckets. Within bucket, column values are randomly permuted to split the linking between different columns. In vertical partition attributes are grouping into columns built on the correlations between the attributes. In each column contains a subset of highly correlated attributes. The idea of slicing is to break the relation between cross columns, but to preserve the relation inside each column. Slicing reduces the dimensionality of the data and preserves good utility than bucketization and generalization.

Because of grouping highly correlated attributes together slicing conserve utility, and conserve the correlations among such attributes. Slicing conserve privacy because it breaks the associations between uncorrelated attributes, which are rare and thus recognizing. When the data set holds quasi-identifiers and one sensitive attributes, bucketization has to break their correlation. Slicing, on the other hand, can group some quasi-identifier attributes with the sensitive attribute, protective attribute correlations with the sensitive attribute. Slicing responsible for privacy protection is that slicing process ensures that for any tuples. There are generally various similar buckets. In slicing partitions attributes into columns. Each column contains a subset of multiple sensitive attributes. Slicing divides tuples into buckets. Each bucket holds a subset of tuples. Inside each bucket, values in each column are randomly permutated for break the linking between different columns. Slicing as a technique for multiple sensitive attributes anonymized published dataset by partitioning the dataset vertically as well as horizontally. Data in which have multiple sensitive attributes used slicing for membership revelation protection and conserves good data useful than generalization and bucketization.

Formalization of Slicing:

Let T be the micro data table to be published. T contains d attributes: A = {A1,A2, . . . ,Ad} and their attribute domains are {D[A1],D[A2], . . . ,D[Ad]}.

A tuple $t \in T$ can be represented as t = (t[A1], t[A2], ..., t[Ad]) where t[Ai] ($1 \leq i \leq d$) is the Ai value of t.

Definition1: (Attribute partition and columns).

An attribute partition include several subsets of A, such that each attribute belongs to exactly one subset. Each subset of attributes is called a column. Specifically, Let there be c columns C1,C2, . . . ,Cc, then $\cup_{i=1}^{c} Ci = A$ and for Any $1 \leq i1 \neq i2 \leq c$, $Ci1 \cap Ci2 = \emptyset$.[1]

For simplicity of discussion, only one sensitive attribute can be considered. If the data contains multiple sensitive attributes, one can either consider them separately or consider their joint distribution [15]. Exactly one of the c columns contains S.

 ISSN: 2321-8169
4261 – 4266

Without loss of generality, let the column that contains S be the last column Cc. This column is also called the sensitive column. All other columns {C1,C2, . . . ,Cc−1} contain only QI attributes.[1]

Definition 2: (Tuple partition and buckets).
A tuple partition consists of several subsets of T, such that each tuple belongs to exactly one subset. Each subset of tuples is called a bucket. Specifically, let there be b Buckets B1,B2, . . . ,Bb, then ∪b i=1Bi = T and for any 1 ≤ i1 6= i2 ≤ b, Bi1 ∩ Bi2 = ∅.[1]

Definition 3: (Slicing).
Given a micro data table T, a slicing of T is given by an attribute partition and a tuple partition. For example, Table I Table II are sliced tables. In Table I, the attribute partition is {{Resting BP},{Hypertension}, {Severity Index}, { Treatment } }. In Table II, the attribute partition is {{Resting BP, Hypertension}, {Severity Index, Treatment}}. Often times, slicing also involves column generalization.[1]

Definition 4: (Column Generalization).
Given a micro data table T and a column Ci = {Ai1,Ai2, . . . ,Aij}, a column generalization for Ci is defined as a set of non-overlapping j-dimensional regions that completely cover D[Ai1] × D[Ai2] × . . . × D[Aij]. A column generalization maps each value of Ci to the region in which the value is contained. Column generalization ensures that one column satisfies the k-anonymity requirement. It is a multidimensional encoding and can be used as an additional step in slicing. Specifically, a general slicing algorithm include the following three Steps: attribute partition, column generalization, and tuple partition. Because each column contains much fewer attributes than the whole table, attribute partition enables slicing to handle high-dimensional data. A key notion of slicing is that of matching buckets.[1]

Definition 5: (Matching Buckets).
Let {C1,C2, . . . ,Cc} be the c columns of a sliced table. Let t be a tuple, and t[Ci] be the Ci value of t. Let B be a Bucket in the sliced table, and B[Ci] be the multi set of Ci Values in B. We say that B is a matching bucket of t iff For all 1 ≤ i ≤ c, t[Ci] ∈ B[Ci].
For example, consider the sliced table shown in Table 1(f), and consider t1 = (22M, 47906, dyspepsia). Then, the set of Matching buckets for t1 is {B1}.[1]

Table I: The Original Data

Name	Resting BP	Hypertension	Severity Index	Treatment	Pulse Rate	Medicines	Education	Profession	salary
Person 1	80	95	4	Kavil	68	Gijamine	ME	Teaching	35000
Person 2	110	85	3	Hypertension	72	Santrifon	MBA	Marketing	20000
Person 3	75	93	2	Flu	87	Oghil	BA	Student	00000
Person 4	40	115	1	Kavil	98	Matrogin	BE	Student	00000
Person 5	105	90	3	Cancer	82	Albhome	MBA	Business	50000

B. Privacy Threats

There are three types of privacy disclosure threats when publishing of micro-data.

a. Membership Disclosure Protection- The first type is membership disclosure, when the data to be published is preferred from a bigger dataset and the selection conditions are sensitive, it is important to prevent an attacker to knowing whether an individual's record is in the data or not.
b. Identity Disclosure Protection- The second type of privacy disclosure thread is identity disclosure, which occurs when an individual is linked to a particular record in the released table. In several situations, one wants to protect against identity disclosure when the attackers is undefined of membership.
c. Attribute Disclosure Protection- The third type is attribute disclosure, occurs when new data about some individuals is published. That means the released data makes it possible to assume the attributes of an individual more correctly than it would be possible before the release. Alike to the case of identity disclosure, required to consider attacker who previously know the membership information. Most of the time Identity disclosure leads to attribute disclosure. Once there is identity disclosure, an individual is re- identified and the equivalent sensitive value is discovered. Attribute disclosure can happen with or without identity disclosure, for example when the sensitive values of all matching tuples are the same.

Table II : The Sliced Table

Resting BP, Hypertension	Severity Index, Treatment	Pulse Rate, Medicines
(80,95)	(4,Kavil)	(68,Gijamine)
(110,85)	(3,Hypertension)	(72,Santrifon)
(75,93)	(2,Flu)	(87,Oghil)
(40,115)	(1,Kavil)	(98,Matrogin)
(105,90)	(3,Cancer)	(82,Albhome)

C. ONE-ATTRIBUTE-PER-COLUMN SLICING DATA:

We examine that while one-attribute-per-column slicing conserve attribute distributional information, each attribute is in its own column therefore it does not protect attribute correlation. In slicing, one group correlated attributes together in one column and preserves their correlation. For

 ISSN: 2321-8169
4261 – 4266

ex., in the sliced table shown in Table correlations between Severity Index and Treatment and correlations between Resting BP and Hypertension are preserved. In fact, the sliced table encodes the same amount of information as the original data with regard to correlations between attributes in the same column.

Table III One Attribute per Column Slicing

Resting BP	Hyperte nsion	Severity Index	Treatment
80	95	4	Kavil
110	85	3	Hypertensi on
75	93	2	Flu
40	115	1	Kavil
105	90	3	Cancer

D. Multiple Column Multiple Attributes

Algorithm:

K-Means Clustering Algorithm:

Clustering in data mining is the process of grouping a set of objects into classes of similar objects [1]. Many clustering algorithms are discussed in the literature and the most important of these are partitioning and hierarchical algorithms. K-means remains one of the most popular clustering algorithms used in practice [3]. The main reasons are it is simple to implement, fairly efficient, results are easy to interpret and it can work under a variety of conditions. The steps to be followed for effective clustering using K-means algorithm are:

Step 1: Begin with a decision on the value of K = number of segments

Step 2: Put any initial partition that classifies the data into K segments. We can arrange the training samples randomly, or systematically as follows:

1) Take the first K training samples as a single-element Segment.

2) Assign each of the remaining (N-K) training samples to the segment with the nearest centroid. Let there be exactly K segments ($C1$, $C2$—CK) and n patterns to be classified such that, each pattern is classified into exactly one segment. After each assignment, re-compute the centroid of the gaining segment.

Step 3: Take each sample in sequence and compute its distance from the centroid of each of the segments. If the sample is not currently in the cluster with the closest centroid switch this sample to that segment and update the centroid of the segment gaining the new sample and cluster losing the sample.

Step 4. Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments. After determining the final value of the K (number of regions) we obtain the estimates the parameters μi, σi and αi for the ith region using the segmented regions.

E. MATHEMATICAL MODEL

1) T = MICRO-DATA TABLE

2) IDENTIFY THE ATTRIBUTES
A = {A1,A2, A3,…AD}

3) D BE THE SET ATTRIBUTE
DOMAIN D = {D[A1], D[A2], …. D[AD]}

4) IDENTIFY THE TUPLE
T = {T[A1], T[A2],…T[AD]G

5) S = SENSITIVE VALUE

6) B = Sliced Bucket
p(t; s) = p(t; B)p(sjt; B)

Where p(t,s)= probability that t takes sensitive value s.

p(t,B)= probability that t is in bucket B.

p(s,B) = probability that t takes sensitive value s given that t in bucket B.
t's column value = t[C1], t[C2]….[Cc]
B's column value = B[C1],
B[C2]….B[Cc]
fi(t, B) = Fraction of occurrences of t(Ci) in B(Ci):
fc(t,B) = Fraction of occurrences of t[Cc - {s}] in B[Cc

Process summary:

1. Extract the data set from the database.

2. Performing anonymization technique on different domains

3. Computes the Overlap sliced table with multiple sensitive attributes on different domains.

4. Attributes are combined and secure data displayed.

F. Experimental Setup

Software Requirement: Basic software specifications are:

H/W System Configuration:-

Processor - Pentium –IV
Speed - 3.0 Ghz
RAM - 256 MB(min)
Hard Disk - 20 GB
Key Board - Standard Windows Keyboard
Mouse - Two or Three Button Mouse

S/W System Configuration:-

Operating System :Windows95/98/2000/XP/Above
Application Server : Tomcat5.0/6.X
Front End : HTML, Java, Jsp
Scripts : JavaScript.
Server side Script : Java Server Pages.
Database Connectivity : Mysql.

IV. Result & Discussion

As mentioned above, limiting attribute to only one column hampers the data utility of the published dataset. The idea of slicing is to release correlated attributes together which then lends to the utility of the anonymized dataset.

 ISSN: 2321-8169
4261 – 4266

Thus, authorizing an attribute to more than one column would release more attribute correlations and thus improve the utility of the released dataset. Table II show the anonymized tables after applying slicing and Table IV show the multiple columns multiple attributes slicing technique. In Table II, Resting BP is grouped with Hypertension and Severity Index is grouped with Treatment. Even if Pulse Rate is grouped with Medicines. In Table IV, the attributes Resting BP, Hypertension and Pulse Rate is existing in one column & Severity Index, Treatment and Medicines is exist in one column & Education, Profession Salary is also existing in one column means they are Multiple Columns Multiple Attributes. This allows highly correlated attributes to group together. This also solves the problem of singular columns by merging correlated attributes into a different column. The associations between uncorrelated attributes are broken; this provides better privacy as the associations between such attributes are less frequent and potentially identifying. The idea of Multiple Columns Multiple Attributes Correlation slicing[1] was suggested by Tiancheng Li. and Ninghui Li. and can be occupied to the attribute partitioning phase of the slicing algorithm.

Table IV Multiple Columns Multiple Attributes

Name	Resting BP, Hypertension, Pulse Rate	Severity Index, Treatment, Medicines	(Education, Profession, Salary)
Person 1	(80,95,68)	(4,Kavil, Gijamine)	(ME,Teaching,35000)
Person 2	(110,85,72)	(3,Hyper, Santrifon)	(MBA,Marketing,20000)
Person 3	(75,93,87)	(2,Flu, Oghil)	(BA,Student,00000)
Person 4	(40,115,98)	(1,Kavil, Matrogin)	(BE,Student,00000)
Person 5	(105,90,82)	(3,Cancer,Albhome)	(MBA,Bussiness,50000)

Multiple column multiple attributes slicing with multiple correlated attributes of multiple columns to protect data from membership discloser. It improve the working efficiency and protection schema other anonymization techniques. Attributes that are highly correlated are in the same column, this preserves the relationships between such attributes. The relations between uncorrelated attributes are damaged; this provides better privacy as the associations between such attributes are less frequent and potentially identifying. Finally the system may test with high dimensional data that show our system work efficiently and provide good result than the traditional systems. In figure 2 shows the graph for anonymization techniques with respect to accuracy of privacy preserving.

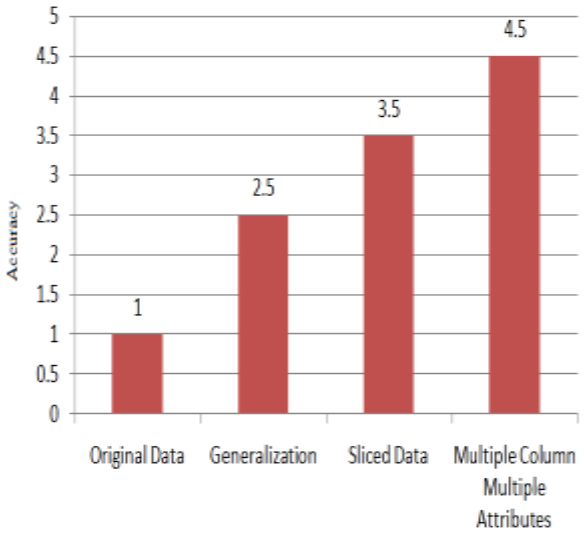

Fig. 1 Graph For Privacy

 ISSN: 2321-8169
4261 – 4266

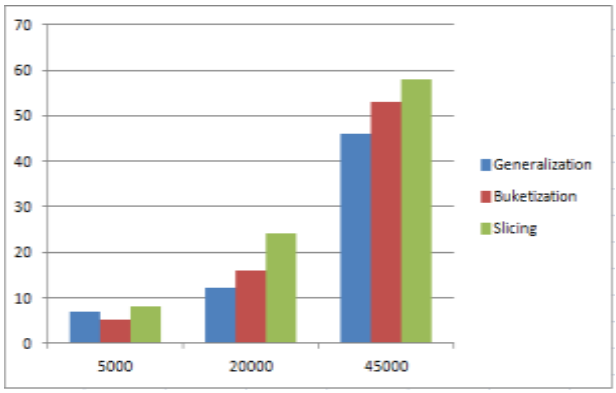

Fig. 2 Graph for cardinality

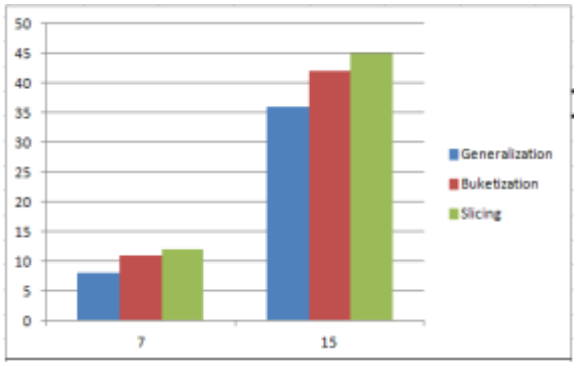

Fig. 3 Graph for Dimensionality

V. CONCLUSION

This paper presents a new technique for increasing the utility of anonymized datasets by improving of slicing. Multiple columns multiple attributes slicing can shows an multiple correlated attributes in single column and this leads to greater data utility because of an increased release of attribute correlations. The Slicing for multiple columns multiple attributes satisfies all the privacy safeguards of traditional slicing such as prevention of attribute disclosure and membership disclosure. Here Slicing for multiple column multiple attributes express the greater data utility provided by improved slicing while satisfying *l*-diversity.

VI. ACKNOWLEDGMENT

I take this opportunity to express my profound gratitude and deep regards to my guide Mr. Yogesh S. Patil for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. I also take this opportunity to express a deep sense of gratitude to my Head of the department Mr. D. D. Patil for her cordial support, valuable information and guidance. Thanks to all those who helped me in completion of this work knowingly or unknowingly like all those researchers, my lecturers and friends.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Senior Member IEEE, Jia Zhang, ember, IEEE, and Ian Molloy Slicing: A New Approach for Privacy Preserving Data Publishing.IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.

[2] C. Aggarwal, On k-Anonymity and the Curse of Dimensionality, Proc. Intl Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.

[3] 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] Brickell and V. Shmatikov, The cost of privacy: destruction of datamining utility in anonymized data publishing In KDD, pages 70-78,2008.

[5] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge, Proc. Intl Conf. Very Data .

[6] M.Terrovitis, N. Mamoulis, and P. Kalnis, Privacypreserving anonymization of set-valued data In VLDB, pages 115125, 2008.

[7] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[8] I. Dinur and K. Nissim, Revealing Information while Preserving Privacy, Proc. ACM Symp. Principles of Database Systems (PODS),pp. 202-210, 2003.

[9] R.C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang, "(_, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 754-759, 2006

[10] Neha V. Mogre, Girish Agarwal, Pragati Patil.A Review On Data Anonymization Technique For Data Publishing Proc. International Journal of Engineering Research Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN:2278-0181

[11] P. Samarati, Protecting Respondents Privacy in Microdata Release, IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[12] L. Sweeney, Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 6, pp. 571-588, 2002.

[13] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern,Worst-Case Background Knowledge for Privacy- Preserv-ing Data Publishing, Proc. IEEE 23rd Int?l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[14] Shyue-Liang Wang,K-anonymity on Sensitive Transaction Items in IEEE International Conferance on Granular Computing 2011.