Overlapping Slicing with New Privacy Model

Suman S. Giri[1]and Mr.Nilav Mukhopadhyay[2]

[1]PG Scholar, Department of Computer Science, Dr. D Y Patil School of engineering and technology, Lohagaon, pune. Maharashtra, India
[2]Associate Professor, Department of Computer Science, Dr. D Y Patil School of engineering and technology, Lohagaon, pune. Maharashtra, India

Abstract- Today, world has growing concern on preserving privacy of census information. There is need of preserving privacy while publishing the data to research center or government agencies. There are various technique have been designed for privacy preserving data publishing such as generalization, bucketization and slicing.

Generalization technique losses considerable amount of information and do not apply for the high dimensional data where as bucketization does not prevent membership disclosure and does not apply for the data that do not have clear separation between quasi identifier and sensitive attribute whereas Slicing releases more attribute correlation and may result in data loss.

In this paper, extension is overlapping slicing which duplicates attribute in more than one column, this releases more attribute correlations. Hence increases privacy and utility of data, by achieving correlation among attributes.

Index Terms- Data Publishing, Data anonymization, Microdata, Privacy preserving, t closeness

I. INTRODUCTION

In recent years wide available personal data has made privacy preserving data mining issue an important one. Privacy is an important factor, while publishing the data to outside world. Many organisations such as Hospitals provides there maintained dataset to reaserch agencies for data analysis.

The data which is going to publish is called microdata which in the form of records. Microdata may contain information about individuals which may include census information such as Disease or salary.

Microdata is mainly divided into three categories such as; 1) explicit identifier: that can clearly identify about an individual, such as social security number, name, address. 2) Quasi identifier: the attribute whose value when taken together can potentially identify an individual such as name and address, name and phone number. 3) sensitive attribute: are the attributes which contains census information such as disease or salary of individuals. Sensitive attribute may provide more knowledge to intruder and may result in information disclosure risk.

II. DATA ANONYMIZATION

Data anonymization is the process of destroying tracks, or the electronic trail, on the data that would lead an eavesdropper to its origins. One of the mechanisms to safeguard personally identifiable information (PII) is to anonymize it. This means removing or obfuscating any identifying information about an individual in a dataset to ensure that it can't be disclosed, while also still allowing valid analysis of the dataset. If data is identifiable when it's collected, then it will still be identifiable when it is stored or analysed unless steps are taken to anonymize it. Anonymization can normally be attempted during collection, retention and disclosure, but any solution will be a balance between anonymity and dataset value, the goal being anonymity with minimal information loss.

There are various privacy models has been developed , such as k-anonymity, l-diversity and t-closeness, which are used with the Data anonymization technique.

Most techniques fall between providing privacy protection and allowing accurate scientific analysis. For example, generalizing an attribute where it's replaced by a less specific value such as age group instead of date of birth is good practice, but limits the level of analysis that can be performed.

III. INFORMATION DISCLOSURE RISK

While publishing very census information about individuals there is information disclosure risk, and data anonymization provides certain level disclosure risk protection. There are mainly three types of disclosure risks as follows:

1. Membership Disclosure
2. Identity Disclosures
3. Attribute Disclosure

Membership disclosure: When the data to be published is selected from a large population and selection criteria is sensitive then it is important to prevent an adversary from learning individuals records is present in database or not

Identity disclosure: The first is when an intruder can assign an identity to any record in the disclosed database. For example, the intruder would be able to determine that record number 7 in the disclosed database belongs to patient Alice Smith. This is called identity disclosure.

Attribute disclosure: Attribute disclosure identification is when an intruder learns something new about a patient in the database without knowing which specific record belongs to that patient. For example, if patients from a particular area in the emergency database had a certain test result, then an intruder does not need to know which record belongs to Alice Smith, if she lives in that particular area then the intruder will discover sensitive information about her. This is called attribute disclosure.

Overlapping slicing is reduces attribute disclosure risk, while achieving attribute disclosure risk there identity disclosure risk can obtain.

IV. T CLOSENESS: A PRIVACY MODEL

Privacy is measured by the information gain of intruder from revealed data. Information gain is nothing but knowledge discovered by the intruder. Before seeing or observing the data, the intruder has prior belief about data and after observing the data the intruder has knowledge about data and it is posterior belief. Information gain is calculated by measuring the difference between prior belief and posterior belief.

Information Gain=prior belief- posterior belief

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

The Earth Mover distance metric is used in order to quantify the distance between the two distributions. Earth Mover distance is used for both numerical and categorical data. Furthermore, the t-closeness approach tends to be more effective than many other privacy preserving data mining methods for the case of numeric attributes.

Example: First an observer has some prior belief *B*0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by **Q**, the distribution of the sensitive attribute value in the whole table, and changes to *B*1. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in and learn the distribution **P** of sensitive attribute values in this class. The observer's belief changes to *B*2. We limit the gain from *B*1 to *B*2 by limiting the distance between **P** and **Q**. Intuitively, if **P** = **Q**, then *B*1 and *B*2 should be the same. If **P** and **Q** are close, then *B*1 and *B*2 should be close as well, even if *B*0 may be very different from both *B*1 and *B*2.

P and **Q** to be close would also limit the amount of useful information that is released, as it limits information about the correlation between quasi identifier attributes and sensitive attributes. However, this is precisely what one needs to limit. If an observer gets too clear a picture of this correlation, then attribute disclosure occurs. The *t* parameter in *t*-closeness enables one to trade off between utility and privacy.

To measure the distance between two probabilistic distributions.
P = (*p*1, *p*2, ..., *pm*),**Q** = (*q*1, *q*2, ..., *qm*), two well-known distance measures are as follows.

1) Variational Distance:

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2}|pi - qi| \quad \ldots\ldots (1)$$

2) And the Kullback-Leibler (KL) distance is defined as:

$$D[P,Q] = \sum_{i=1}^{m} p_i \log \frac{P_i}{q_i} = \text{H(P)- H(P,Q)} \quad \ldots\ldots (2)$$

Where $H(P) = \sum_{i=1}^{m} p_i \log p_i$ is the entropy of P and $H(P,Q) = \sum_{i=1}^{m} p_i \log q_i$ is the cross entropy of P and Q

EMD(Earth Movers Distance)

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance. EMD can be formally defined using the well-studied transportation problem.

Let **P** = (p1, p2, ...pm),**Q** = (q1, q2, ...qm), and dij be the ground distance between element i of **P** and element j of **Q**. We want to find a flow F = [fij] where fij is the flow of mass from element i of **P** to element j of **Q** that minimizes the overall work:

$$\text{Work(P,Q,F)} = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij} \quad \ldots\ldots (3)$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m,\ 1 \leq j \leq m \quad \ldots (i)$$

$$p_i - \sum_{j=1}^{m} fij_0 + \sum_{j=1}^{m} fij = q_i \qquad 1 \leq I \leq m \ldots (ii)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} fij = \sum_{i=1}^{m} pi = \sum_{i=1}^{m} qi = 1 \quad \ldots (iii)$$

These three constraints guarantee that **P** is transformed to **Q** by the mass flow F. Once the transportation problem is solved, the EMD is defined to be the total work,i.e.,

$$\text{D[P,Q] = WORK (P,Q,F)} = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij} \quad \ldots\ldots (4)$$

Earth mover distance is used to calculate the distance between two distributions. EMD is used for both numerical data and the categorical data.

EMD for Numerical Data: Numerical data are in order that's why order distance is calculated for the numerical data. For categorical data, equal distance and hierarchical distance is need to calculate.

Let the attribute domains are $\{d_1, d_2, \ldots, d_m\}$ where d_i is the i_{th} smallest value. For numerical data the order distance between two values is calculated by number of values between them

$$\text{Order list } (v_i, v_j) = |i-j|/(m-1) \quad \ldots\ldots (5)$$

Ordered distance is measured by metrics. It is nonnegative and use triangle inequality and symmetry property. To calculate the ordered distance there is need to consider flows that transport distribution mass between adjacent elements, because any transportation between two more distance need to consider flows that transport distribution mass between adjacent elements, because any transportation between two more distance distant elements can be equivalently decomposed into several transportations between adjacent elements. Based on this observation, minimal work can be achieved by satisfying all elements of Q sequentially

EMD for Categorical Data: For categorical attribute we need to consider two distance measure, first is Equal distance which is ground distance between any two categorical attribute is defined to be 1. It iseasy to verify that this is a metric. As the distance between any two values is 1, for each point that $pi - qi > 0$, one just needs to move the extra to some other points. Thus we have the following formula:

$$D[P,Q] = \frac{1}{2}\sum_{i=1}^{m}|pi - qi| = \sum_{p_i \geq q_i}(pi - qi) = -\sum_{p_i < q_i}(p_i - q_i) \quad \ldots . (6)$$

V. LITERATURE SURVEY

L. Sweeney.,(2002) has proved that elimination of sensitive attributes from microdata is not sufficient to preserving privacy. There are a few solutions proposed in the literature to protect against the information linkage.

L. Swenny and samarati has proposed privacy model called K anonymity which is used with generalization. K anonymity require that each record should be indistinguishable at least k-1 record from other record. K anonymity was the first privacy model used to anonymize data. K anonymity protects against identity disclosure but does not work with attribute disclosure. K anonymity has problem against homogeneity attack and background knowledge attack.

A.Macchanavajjabala has introduced L diversity in 2006, In which each equivalence class has at least l well-represented sensitive values. L diversity is used in 2007 by N. Koudas, D. Shrivastava, and used with bucketization and slicing. Bucketization does not protect for membership disclosure risk and it doesn't not differentiate between quasi identifier and sensitive attribute.

T. Li and N. Li. In 2009 has emerged new approach i.e. the tradeoff between privacy and utility in data publishing.

In 2012 slicing technique has been proposed for data anonymization which works for high dimensional data, and also protect from membership disclosure risk. Due to high attribute correlation privacy violation may happen in slicing techniques . Data slicing can also be used to prevent membership disclosure and is efficient for high dimensional data and preserves better data utility. T closeness a new privacy measure is proposed by N. Li in 2007. In 2007 N. Li , T Li has proved that t closeness can be used with anonymization techniques. k-anonymity prevents identity disclosure but not attribute disclosure To solve that problem l-diversity requires that each eq. class has at least l values for each sensitive attribute But l-diversity has some limitations t-closeness requires that the distribution of a sensitive attribute in any equivalent class is close to the distribution of a sensitive attribute in the overall table.

VI. OVERLAPPING SLICING

Problem statement: Privacy preserving data publishing is an issue now days**.** While data get published to any agencies, there is risk of information disclosure. While reducing information disclosure risk there is loss of data utility. Slicing may fail to achieve data privacy and utility because during attribute partitioning sensitive attribute is grouped into single column Hence there is less correlation between attributes, and l diversity may does not work for attribute disclosure risk.

Proposed technique: The proposed technique is overlapping slicing in which attributes are duplicated in more than one column and easy to achieve more correlation between attribute. Overlapping slicing partitions attribute both horizontally and vertically. In vertical partitioning more correlated attributed are taken into one group and uncorrelated attributed are grouped separately. In horizontal partitioning tuple are grouped to form buckets, after grouping tuples values of column are randomly permuted. Overlapping slicing works in three main steps:

1. Attribute partitioning
2. Tuple partitioning
3. Column generalization

Attribute partitioning : In attribute partitioning, correlation of the attribute are measured to form there group. To measure the correlation mean square contingency coefficient is used. Mean square coefficient is achieved by following formula:

$$\emptyset^2(A_1, A_2) = \frac{1}{\min\{d_1,d_2\}-1}\sum_{i=1}^{d1}\sum_{j=1}^{d2}\frac{(f_{ij}-f_i f_j)^2}{f_i f_j} \quad \ldots\ldots .(7)$$

Given two attributes A1 and A2 with domains $\{v_{11}, v_{12}, ..., v_{1d1}\}$ and $\{v_{21}, v_{22}, ..., v_{2d2}\}$, respectively. Their domain sizes are thus d1 and d2, respectively. The mean-square contingency coefficient between A1 and A2 is defined as:

Here, $f_{i\cdot}$ and $f_{\cdot j}$ are the fraction of occurrences of v_{1i} and v_{2j} in the data, respectively. f_{ij} is the fraction of co-occurrences of v_{1i} and v_{2j} in the data. Therefore, $f_{i\cdot}$ and $f_{\cdot j}$ are the marginal totals of f_{ij} : $f_{i\cdot} = \sum_{j=1}^{d=2} fij$ and $f_{\cdot j} = \sum_{i=1}^{d=1} fij$. It can be shown that $0 \leq \varphi^2(A1,A2) \leq 1$.

Attribute clustering: Having computed the correlations for each pair of attributes, we use clustering to partition attributes into columns. We use k mediod for clustering. In algorithm each attributes is taken as point in clustering space. The distance between two attributes in the clustering space is defined as $d(A1,A2) = 1 - \varphi 2(A1,A2)$, which is in between of 0 and 1. Partition around k mediod algorithm is used for clustering.

Algorithm Partitioning Around Medoid (PAM)

Initialize: randomly select *k* of the *n* data points as the medoid

1. Associate each data point to the closest medoid.
2. For each medoid *m*
1. For each non-medoid data point *o*
1. Swap *m* and *o* and compute the total cost of the configuration
3. Select the configuration with the lowest cost
4. Repeat steps 2 to 4 until there is no change in the medoid.

Tuple partitioning: In this step tuples are grouped to form bucket. Mondrian algorithm is used for tuple partitioning.

Algorithm tuple-partition(T, t)

1. Q = {T}; SB = $\emptyset$.

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 t closeness-check(T, Q ∪ {B1,B2} ∪ SB, t)
6. Q = Q ∪ {B1,B2}.
7. else SB = SB ∪ {B}.
8. return SB

Figure 1 gives the description of the tuple-partition algorithm. The algorithm maintains two data structures: (1) a queue of buckets Q and (2) a set of sliced buckets SB. Initially, Q contains only one bucket which includes all tuples and SB is empty (line 1). In each iteration (line 2 to line 7), the algorithm removes a bucket from Q and splits the bucket into two buckets (the split criteria is described in Mondrian [17]). If the sliced table after the split satisfies ℓ-diversity (line 5), then the algorithm puts the two buckets at the end of the queue Q (for more splits, line 6). Otherwise, we cannot split the bucket anymore and the algorithm puts the bucket into SB (line 7). When Q becomes empty, we have computed the sliced table. The set of sliced buckets is SB (line 8).

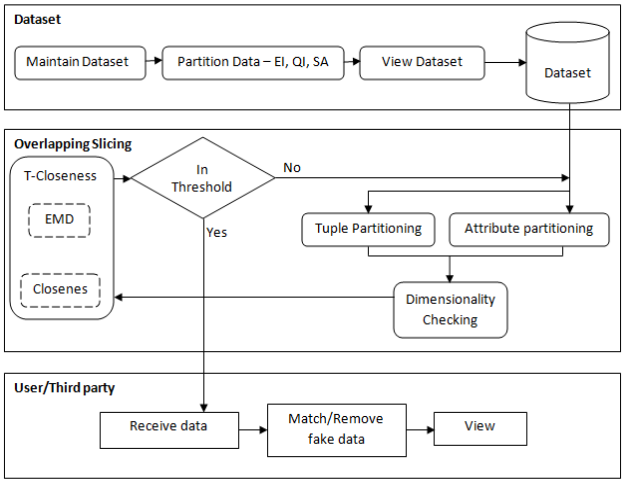

Fig1 Block diagram for overlapping slicing

Architecture of Overlapping slicing:

Architecture of the overlapping slicing as shown in figure. In first phase maintain dataset in the form of tables and records. Attributes are classified into three main categories i.e. Explicit identifier, quasi identifier and sensitive attributes.

In second step overlapping slicing is performed on the data set i.e. vertical and horizontal portioning is performed on the dataset. During tuple partioning t closeness privacy model is used to check achieved privacy of data. T closeness calculates the EMD of data to achieve privacy.After performing partitioning dimensionality of data is checked by using dimensionality check algorithm. After completion of overlapping slicing , we provide fake tuples to hide the original tuple.

VII. EXPECTED RESULT

We evaluate the effectiveness of overlapping slicing in preserving data utility and protecting against attribute disclosure, identity disclosure and membership disclosure as compared to generalization, bucketization and slicing.

Data set

Some preprocessing steps must be applied on the anonymized data before it can be used for workload tasks. First, the anonymized table computed through generalization contains generalized values, which need to be transformed to some form that can be understood by the classification algorithm. Second, the anonymized table computed by bucketization or slicing contains multiple columns, the linking between which is broken. We need to process such data before workload experiments can run on the data.

Handling generalized values: In this step, we map the generalized values (set/interval) to data points. the Mondrian algorithm assumes a total order on the domain values of each attribute and each generalized value is a sub-sequence of the total-ordered domain values. There are several approaches to handle generalized values. The first approach is to replace a generalized value with the mean value of the generalized set. For example, the class 9th, 10th and 11th replaced by 10th. The second approach is to replace a generalized value by its lower bound and upper bound. In this approach, each attribute is replaced by two attributes, doubling the total number of attributes. For example, the Education attribute is replaced by two attributes Lower-Education and Upper Education; for the generalized Education level {9th, 10th,11th}, the Lower-Education value would be 9th and the Upper-Education value would be 11th. We use the second approach in our experiments. Handling bucketized/sliced data. In both bucketization and slicing, attributes are partitioned into two or more columns. For a bucket that contains k tuples and c columns, we generate k tuples as follows. We first randomly permutated the values in each column. Then, we generate the ith ($1 \leq i \leq k$) tuple by linking the i-th value in each column. We apply this procedure to all buckets and generate all of the tuples from the bucketized/sliced table. This procedure generates the linking between the two columns in a random fashion.

Table 1 contains the record of original microdata table in which Name is explicite identifier which removed in first step. Age, gender, and zipcode are quasi identifier and remaining two diseases and occupation are the sensitive attribute.

Table 2 is overlapped slicing table in which explicit identifier Name is removed from table and qusi identifier are grouped together with one sensitive attribute and another group of sensitive attribute. The value of sensitive attribute is randomly permuted to achieve more privacy. Sensitive attributes are partitioned with both attribute therefore more attribute correlation is achieved and utility of data is increased.

TABLE I
ORIGINAL MICRODATA TABLE

Name	Gender	Age	Zipcode	Disease	Occupation
A	M	22	410505	FLU	Student
D	F	22	410905	FLU	Student
E	F	35	410702	Bronchitis	Service
N	F	50	410208	Cancer	Retire
Y	M	59	410507	Bronchitis	Business
Z	M	67	410906	Cancer	Retire
P	M	62	410305	BP	Business
H	F	63	410308	BP	business

Expected Result Set

TABLE II
OVERLAPPED SLICED TABLE(PROPOSED SYSTEM)

(Age, gender, Disease)	(Zipcode, Disease, occupation)
22,M,flu	410505,flu,Student
22,F,flu	410905, flu, Student
35, F, bronchitis	410702,bronchitis, Service
50,F,cancer	410208,cancer, Retire
59, M, bronchitis	410507,bronchitis, Business
67, M, cancer	410906, cancer, Retire
62, M, BP	410305, BP, Business
63, F, BP	410308, BP, business

VIII. CONCLUSION

Anonymization technique is powerful method for privacy preserving of published data. This paper presents a new anonymization method that is overlapping slicing with new privacy model i.e. t closeness for privacy preserving and data publishing. This method overcomes the limitations of slicing and preserves better utility while protecting against privacy threats. Overlapping slicing that how slicing is used to prevent attribute disclosures.

The general methodology of this work is before data anonymization one can analyze the data characteristics in data anonymization. The basic idea is one can easily design better anonymization techniques when we know the data perfectly. Finally, we have some advantages of overlapping slicing comparing with generalization and bucketization and slicing. Overlapping slicing is a promising technique for handling high dimensional data. By increasing the correlation among data privacy is preserved.

ACKNOWLEDGEMENT

The authors would like to thank Tiancheng Li, Ninghui, Jiang Zhang, Ian Molloy for their helpful divotions, and they thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, Slicing: A New Approach to Privacy Preserving Data Publishing, IEEE 2012 Transactions on Knowledge and Data Engineering, volume:24,Issue:3

[2] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and _-Diversity, Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference

[3] C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, pages 901–909, 2005.

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

[5] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB, pages 139–150, 2006.

[6] L. Sweeney. k-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzz., 10(5):557–570, 2002.

[7] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz., 10(6):571–588, 2002

[8] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In KDD, pages 517–526, 2009.

[9] Balamurugan Shanmugam, Visalakshi Palanisamy, Modified Partitioning Algorithm for Privacy Preservation in Microdata Publishing with Full Functional Dependencies, Australian Journal of Basic and Applied Sciences, 7(8): 316-323, 2013 ISSN 1991-8178

[10] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala, Privacy-Preserving Data Publishing, Foundations and TrendsR_ in Databases Vol. 2, Nos. 1–2 (2009) 1–167

AUTHORS

First Author – Suman S. Giri, PG Scholar, Department of Computer Science, Dr. D Y Patil School of engineering and technology, Lohagaon, pune. Maharashtra, India email:suman.giri311@gmail.com
Second Author – Mr.Nilav Mukhopadhyay, Associate Professor, Department of Computer Science, Dr. D Y Patil School of engineering and technology, Lohagaon, pune. Maharashtra, India e-mail: nilove18@gmail.com