ISSN 2229-5518

Multi-Privacy Collaborative Data publishing with Efficient Anonymization Techniques

K.S.Gangatharan, M.S.Thanabal

Abstract— Privacy-preserving in collaborative data publishing provides methods and tools for publishing the data while protecting the sensitive information in the data set. The success of data mining in privacy relies on the information sharing and quality of data in a distributed environment. Several anonymization techniques have been proposed such as bucketization, generalization which does not prevent membership disclosure and results in loss of information. Slicing used for high dimensional data and prevent membership disclosure but it takes place at only one column. The above techniques where each of them has taken the solution for creating a privacy when the microdata publishing. The system proposed in this paper tells the m-privacy and m-adversary technique with overlapping slicing concept which takes place more than two attribute column. M-privacy protects anonymized data from adversaries for a given privacy constraint. This technique shows the better utility and efficiency than the previous techniques.

Index Terms— adversary, anonymization, microdata, m-privacy, privacy-preserving, slicing, security.

1 INTRODUCTION

Data sharing is needed nowadays and there is a growth in the sharing of information all around. Data mining is the process of extracting large amount of hidden useful data from the database while preserving individual privacy. From the author in [4], Data mining is also said to be Knowledge Discovery in Database (KDD) the process to extract or discover knowledge and summarizes data from different database. Nowadays data mining is used as many forms with a strong consumer focus in commercial purpose and marketing organizations. In [1] the healthcare domain, a national agenda the Nationwide Health Information Network is formed to share information over the internet among hospitals and other providers, with privacy protection.

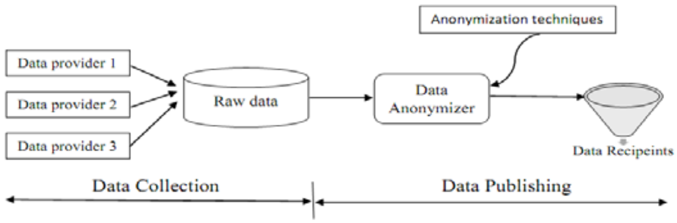
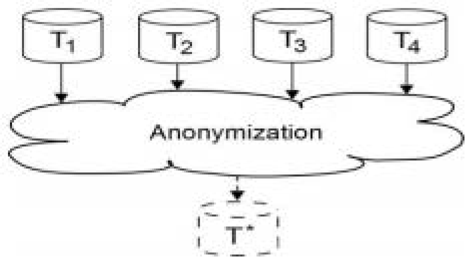


Fig. 1. System Process

Privacy preserving data and data publishing method have received attention as promising approach for information sharing while preserving individual privacy. In a distributed manner the data from multiple data providers has two main process used for anonymization. The author in [1] states two approach, First approach is to anonymize the data independently from each provider (anonymize-and-aggregate) which results in potential loss of data utility. Second approach is collaborative data publishing is to anonymize the data from all individuals which come from one source (aggregate and anonymize), using by trusted third-party (TTP) or by the Secure Multi-party Computation (SMC) protocols. In the data collection model, the raw data collects data from individuals. In the data publishing model, the data holder releases data to the public called the data recipient, then process data mining technique.

1.1 Collaborative data publishing

The goal of collaborative publishing is to publish integrated data T*, in an anonymized view. Attacks may be from one or more group of internal and external entities that reduce privacy of data using background knowledge. Secure Multi-Party Computation (SMC) or Trusted third party protocols (TTP) are used to successfully carried out the Collaborative data publishing that guarantees the information of a particular individual is protected. With help of SMC the privacy is maintained and results better data utility.



- *K.S.Gangatharan is currently pursuing master degree program in Computer Science Engineering Department in PSNA College of Engineering and Technology, Dindigul, India.*
  E-mail: ganga.ks@gmail.com
- *M.S.Thanabal is currently working as an Associate Professer in Computer Science Engineering Department in PSNA College of Engineering and Technology, Dindigul, India.*
  E-mail: msthanabal@psnacet.edu.in

Fig. 2. Collaborative Data Publishing

1.2 Data Anonymization

Data Anonymization is a technique that convert normal text data into a non-readable form and remove traces from the source. Data anonymization technique in privacy-preserving collaborative data publishing has become an important nowadays for secure publishing. Microdata contains information about a person or an organization is to preserve from adversary. Most popular approaches in anonymization techniques are Generalization, Bucketization and slicing. The author from [2] categorize the number of attributes in each record as 1) Identifiers such as Name, Number etc., are the key attributes that uniquely identify the individuals and removed before it release. 2) Some attributes are Sensitive Attributes (SAs) such as medical records and salary etc., 3) some are may be Quasi-Identifiers (QI) such as birth date, Zipcode, and gender etc., which can potentially identify an individual and used to link anonymized dataset with other datasets.

2 Principles used in privacy preserving

Various principles are used in privacy-preserving data mining for collaborative data publishing such as l-diversity, k-anonymity, t-closeness etc. And various protocols are used for securing data from adversary.

2.1 K-anonymity

K-anonymity from [7] [8] stats that each tuple in the published database table not differentiate from at least k-1 of other tuples. The concept in k-anonymity is to minimize the granularity representation of the microdata in such a way that a given tuple records cannot be separable from k-1 of other tuple records. To prevent record linkage through quasi identifier, the record in the one table has some value of QID, at least k-1 of other records also have the value quasi identifier or otherwise minimize the size of group on QID. A table which satisfying this minimum requirement is called k-anonymous. In all cases anonymity cannot be safeguard against attribute disclosure. From the author in [3] addressed two forms of attack in k-anonymity.

Homogeneity type Attack:

When the attacker known the non-sensitive information from individual then sensitive information may be identified based on the known information.

Background Knowledge type Attack:

In this type of attack, the adversary can use a correlation between one or more QID attributes with the sensitive attribute in order to retrieve possible values of the sensitive information further.

2.2 L-diversity

K-anonymity is intend to prevent identification of a tuple record and it may not always be intend to prevent inference of the sensitive attribute values of the record. To overcome the limitations, the l-diversity was proposed in [3] which focus on maintaining diversity of the sensitive attributes but also maintains the minimum group size of k. It put constraints on minimum number of distinct value attributes for any sensitive attributes in an equivalence class. From [5] l-diversity has equivalence class if there is one or more well-known represented values for the sensitive attributes. It must ensure that the important attribute value of a particular individual cannot be identified unless the adversary has enough background knowledge. The author from [9] addressed two forms attack in l-diversity.

Skewness type attack:

While the attribute distribution is skewed, the diversity does not prevent attribute disclosure.

Similarity type attack:

Sensitive information leakage occurs when ℓ-diversity requirement assures “diversity” in each group of sensitive values does not consider into account.

3 Related work

3.1 Generalization

Generalization technique is one of the commonly used anonymized approaches, which replaces quasi-identifier attribute values with values that are less specified but semantically consistent. In a group, all quasi-identifier values would be generalized to the entire QID space. The author in [2] stats if any two transactions have distinct values in a group for a certain column then all information about that individual item in the group is lost. The quasi identifier taken in the process includes all possible tuples in the system log. Presence high-dimensionality of the quasi-identifier, generalization method would inhibit information loss and rendering the data unworthy. In order to avoid this drawback, generalization, to keep the records close to each other in the same bucket so that the records would not loss information and results generalization to be effective. For high-dimensional data, distance between the data points are similar with each other.

Drawback

(1) It loss on high dimensional data.
(2) Due to uniform distribution nature it results too much information loss.

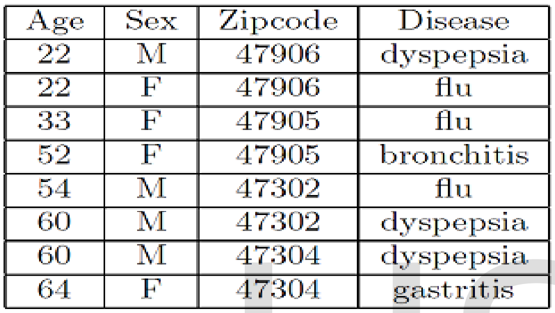
3.2 Bucketization

The term bucketization is to partition the tuples into buckets. By random permuting it separate sensitive attribute from non-sensitive attribute within each bucket. Then the data in the bucket consists of sensitive attribute values. From [2] the notion of bucketization is to partitioning the tuples into buckets by horizontal partition and apply a random permutation to the column in the bucket contains sensitive values, then the resulting set is published. The anonymized data consist of sensitive attribute values in buckets. From the author in [8], when to anonymize high-dimensional data bucketization has been used. While compare to generalization it has better data.

Drawback

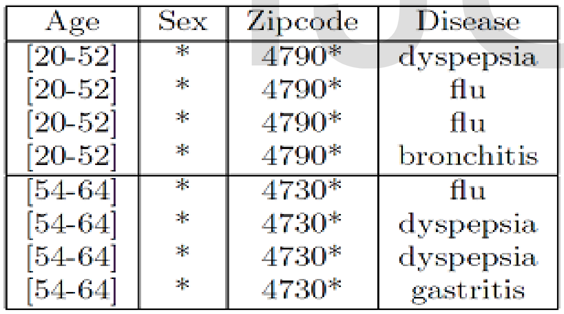
(1) It does not prevent membership disclosure.
(2) It requires a clear separation of sensitive attributes and quasi identifier.
(3) It breaks the attribute correlation between sensitive attributes and quasi identifier.

Table 1
Original table



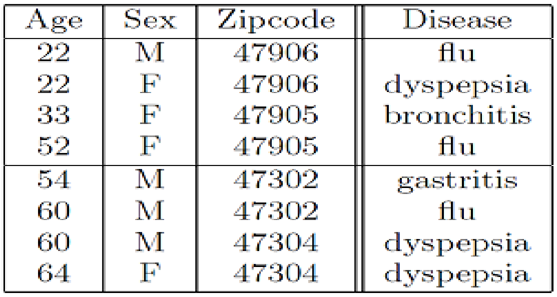
Age	Sex	Zipcode	Disease
22	M	47906	dyspepsia
22	F	47906	flu
33	F	47905	flu
52	F	47905	bronchitis
54	M	47302	flu
60	M	47302	dyspepsia
60	M	47304	dyspepsia
64	F	47304	gastritis

Table 2
Generalization table



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

Table 3
Bucketization table



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

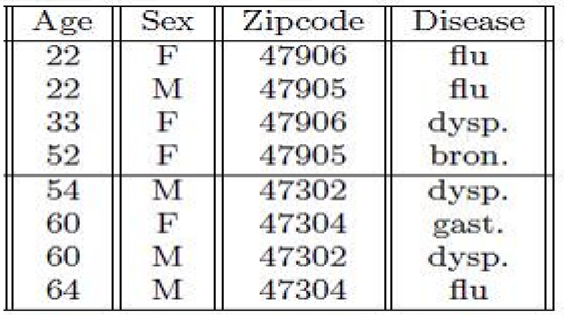
3.3 Slicing

Techniques like bucketization, generalization placed to preserve privacy but they exhibit attribute disclosure. Slicing technique is used to overcome this drawback of generalization and bucketization while handling high dimensional data. From [2] slicing makes the dataset partition both horizontally and vertically. Grouping of attributes into columns is done by vertical partitioning from the correlation of attributes. Each columns are highly correlated with subset of attributes. Grouping of tuples into buckets done by horizontal partitioning. Each column values are randomly sorted in the bucket to break the link among different columns. The scope of slicing is to break the association among columns and also preserve association between columns. This minimize the size of the data and preserves better utility. It preserves utility by grouping highly-correlated attributes, and preserves the correlations among attributes. The table 4 contains four columns, where each of the column contains exactly only one attribute. Sliced table shown in table 5 contains two columns, the first column has {Age, Sex} and the second column has {Zipcode, Disease}. Both table contain two buckets and each of them has four tuples. The values in the bucket from each column is randomly split to break the link between different columns.

Draw back

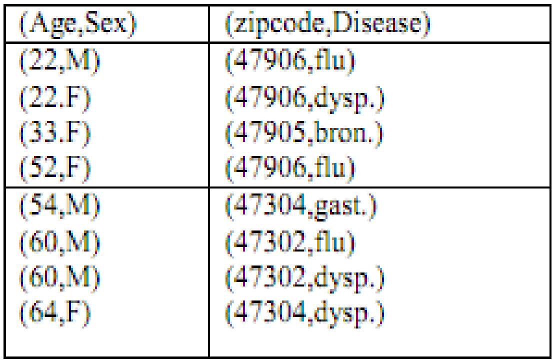
(1) It takes place where an attribute present exactly at one column.
(2) The extension slicing techniques duplicates an attribute in more than one columns.

Table 4
One-attribute-per-column slicing



Age	Sex	Zipcode	Disease
22	F	47906	flu
22	M	47905	flu
33	F	47906	dysp.
52	F	47905	bron.
54	M	47302	dysp.
60	F	47304	gast.
60	M	47302	dysp.
64	M	47304	flu

Table 5
Two-attribute-per-column slicing



(Age,Sex)	(zipcode,Disease)
(22,M)	(47906,flu)
(22.F)	(47906,dysp.)
(33.F)	(47905,bron.)
(52,F)	(47906,flu)
(54,M)	(47304,gast.)
(60,M)	(47302,flu)
(60,M)	(47302,dysp.)
(64,F)	(47304,dysp.)

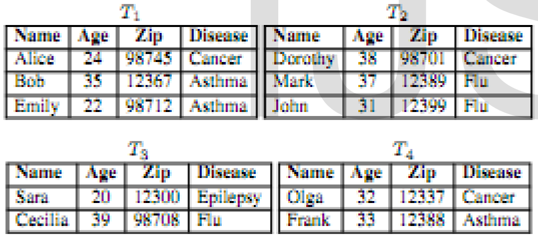
4 Design of Anonymization Techniques

In this work, m-privacy addresses the new threat from m-adversaries and make some contributions for privacy-preserving. The m-privacy technique models the data knowledge of an m-adversary and secure anonymized data against from such adversaries for a given privacy constraint. For example, the records in each equivalence group still satisfy l-diversity for any m-adversary while anonymization satisfy m-privacy. To verify m-privacy efficiently for a given a set of records heuristic algorithms is used. It utilizes effective pruning strategy for an equivalence group monotonicity property to the given privacy constraints. For checking m-privacy it uses data provider-aware anonymization algorithm to ensure high data utility and efficiency. In an overlapping strategy, combines two well-known anonymization techniques while ensuring data with high utility.

4.1 m-privacy

In common, multiple data providers may combine with one another, hence having access to their data, or an individual may have access to various databases. Here the data provider identify the new type of attack called "insider attack". And the data which may access outside by the attacker is said to be "outside attacker".

Table 6
Data providers



T_1

Name	Age	Zip	Disease
Alice	24	98745	Cancer
Bob	35	12367	Asthma
Emily	22	98712	Asthma

T_2

Name	Age	Zip	Disease
Dorothy	38	98701	Cancer
Mark	37	12389	Flu
John	31	12399	Flu

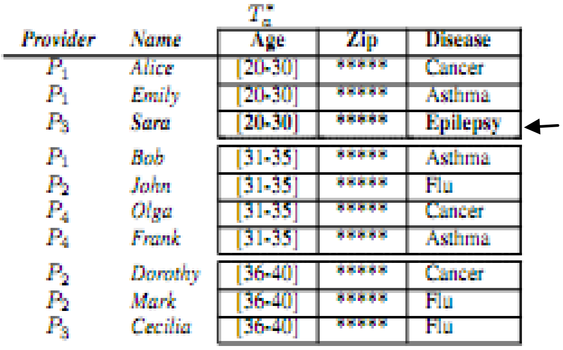
T_3

Name	Age	Zip	Disease
Sara	20	12300	Epilepsy
Cecilia	39	98708	Flu

T_4

Name	Age	Zip	Disease
Olga	32	12337	Cancer
Frank	33	12388	Asthma

Table 7
M-adversary

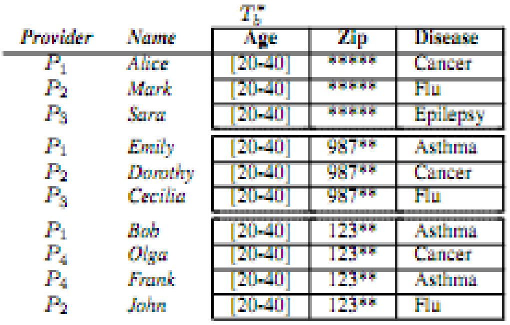


T_a^*

Provider	Name	Age	Zip	Disease
P_1	Alice	[20-30]	*****	Cancer
P_1	Emily	[20-30]	*****	Asthma
P_3	Sara	[20-30]	*****	Epilepsy ←
P_1	Bob	[31-35]	*****	Asthma
P_2	John	[31-35]	*****	Flu
P_4	Olga	[31-35]	*****	Cancer
P_4	Frank	[31-35]	*****	Asthma
P_2	Dorothy	[36-40]	*****	Cancer
P_2	Mark	[36-40]	*****	Flu
P_3	Cecilia	[36-40]	*****	Flu

For example, assume that hospitals has four data providers P_1, P_2, P_3, and P_4 try to anonymize their individual patient databases T_1, T_2, T_3, and T_4 collaboratively. In each transaction database, Name as identifier, (Age, Zip) as quasi-identifier, and Disease as sensitive attribute. T^*_a one possible quasi-identifier group-based anonymization technique that assures k-anonymity (k = 3) and l-diversity (l = 2). Taken that l-diversity contain individual records with at least one different sensitive attribute values in each equivalence group. When a hacker from the hospital P_1 has access to T_1, may extract all records from T∗a then also from T_1 and points out that there is only one individual between 20 to 30 ages. Combining this information with known background knowledge, provider P_1 can identify sensitive attribute value from T_3 Sara's record (pointed in the table 7) and her disease Epilepsy.

Table 8
M-privacy



T_b^*

Provider	Name	Age	Zip	Disease
P_1	Alice	[20-40]	*****	Cancer
P_2	Mark	[20-40]	*****	Flu
P_3	Sara	[20-40]	*****	Epilepsy
P_1	Emily	[20-40]	987**	Asthma
P_2	Dorothy	[20-40]	987**	Cancer
P_3	Cecilia	[20-40]	987**	Flu
P_1	Bob	[20-40]	123**	Asthma
P_4	Olga	[20-40]	123**	Cancer
P_4	Frank	[20-40]	123**	Asthma
P_2	John	[20-40]	123**	Flu

In the example table 8, T∗b is an anonymized data that satisfies m-privacy (m = 1) w.r.t. K-anonymity (k = 3) and l-diversity (l = 2). From table 7, T∗a satisfies 0-privacy with respect to constraint C = k-anonymity (k = 3) and l-diversity (l = 2), while T∗b satisfies 1-privacy with respect to the same constraint C.

Algorithm 1: m-Privacy Verification Algorithm

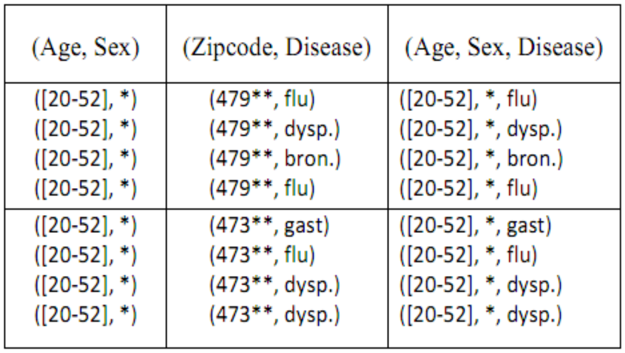
```
Data: Anonymized records T∗ from providers P, and the m.
Result: true if T∗ is m-private w.r.t. C, false otherwise.
1. For each i = 0, 1… m do
2. For each I ∈ ordered coalitions of size (i) do
3. $T_I = \cup_{Pj \in I}$ records of (Pj)
4. For each $S \in 2^{T_I \wedge}$ providers of (S) = I do
5. If privacy does not hold for (T∗ \ S) then
6. Return false
7. Return true
```

For each cardinality of m-adversary, it generates all possible combination of adversaries (line 1 to 2). Then, for each collusion it generates all possible subsets of the records such that each data provider participates to this set with at least one record (line 4). Finally, it verifies if such subsets can be used in attacks to breach privacy (line 5). If the attack is successful, then no further checks are necessary, and the algorithm returns negative answer (line 6). After verifying that all possible subsets of records provided by any m-adversary are not enough to breach privacy, the algorithm returns positive answer.

4.2 Overlapping technique

The author in [10] stats that overlapping slicing can be done in three steps Attribute type partitioning, Tuple type partitioning and Column type generalization it shown in the table IV and V. By taking this idea, overlapping is obtained from the m-privacy table VIII with respect to three attribute per-column slicing technique. It combines three attributes age, sex, and disease in the single column. It ensures data with high utility then the existing approaches.

Table 9
Overlapped table



(Age, Sex)	(Zipcode, Disease)	(Age, Sex, Disease)
([20-52], *)	(479**, flu)	([20-52], *, flu)
([20-52], *)	(479**, dysp.)	([20-52], *, dysp.)
([20-52], *)	(479**, bron.)	([20-52], *, bron.)
([20-52], *)	(479**, flu)	([20-52], *, flu)
([20-52], *)	(473**, gast)	([20-52], *, gast)
([20-52], *)	(473**, flu)	([20-52], *, flu)
([20-52], *)	(473**, dysp.)	([20-52], *, dysp.)
([20-52], *)	(473**, dysp.)	([20-52], *, dysp.)

5 Experiment Results

The microdata values from the table is taken and retrieved from the database. Before the experiments done on the data, preprocessing must be applied to the table. The sensitive attribute and quasi identifier are examined after preprocessing computed. Then the anonymization techniques is carried out on the microdata. In order to evaluate the performance level of anonymization techniques the accuracy can be measured against the several privacy threats such as identity, membership and attribute disclosures. The accuracy of the process can be determined by matching of fake tuples and buckets to the original microdata. The experiments illustrate that m-privacy and overlapping technique preserves better data utility. The results from the m-privacy and overlapping techniques shows better performance than the existing techniques such as generalization, bucketization, slicing.

6 Conclusion

This paper focus on m-privacy with overlapped slicing technique for collaborative data publishing in the distributed environment. It takes advantage of overlapping slicing method with m-privacy technique to protect attacks from inside. It results in secure data publishing in a high dimensional data. By having multiple data providers this technique shows better data utility than the previous anonymization techniques.

References

[1] 1. Slawomir Goryczka, C Benjamin, Li Xiong, M. Fung, “m-privacy for Collaborative Data Publishing” {IEEE} Trans. Knowledge Data Eng., volume: pp no:99 year 2013.

[2] 2. Tiancheng Li, Ninghui Li, Ian Molloy and Jian Zhang, “slicing: a new approach for privacy preserving data publishing”, {IEEE} Trans. Knowledge Data Eng., vol.23, no.2, 2011.

[3] 3. M. Karthikeyan, P. Mayil Vel Kumar, “l-diversity on k-anonymity with External Database for improving Privacy Preserving Data Publishing” , Int’l Journal of Computer Applications, Volume 54, No.14, September 2012.

[4] 4. Gonzalo mariscal, Oscar marban, Covadonga Fernandez, “A survey of data mining and knowledge discovery process models and methodologies”, The Knowledge Engineering Review, Vol. 25:2, 137–166.

[5] 5. Sowmyarani C N, G N Srinivasan, “Survey on Recent Developments in Privacy Preserving Models”, IJCA (0975-8887) Volume 38, No.9, Jan 2012.

[6] 6. C. Clifton, W. Jiang, “secure distributed framework for achieving k-anonymity”, vldbj Volume 15, No.4, 0316–0333, 2006.

[7] 7. Pierangela Samarati, “protecting respondent’s identities in microdata releases” {IEEE} Trans. Knowledge Data Eng., Volume 13, No.6, November 2001.

[8] 8. L. Sweeney, “k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[9] 9. Ninghui, Tiancheng, Venkatasubramanian, “t-Closeness: Privacy beyond k-Anonymity and l-Diversity”, IEEE 23rd International Conference on Data Engineering, ICDE, April 2007.

[10] 10. V. Poovarasi, D.Vijay Anand, “Overlapping Slicing: An Approach to Privacy Preserving Data Publishing”, ijarcsa Volume 1, Issue No. 2, April 2013.