An Efficient Technique to Secure data Access for Multiple Domains using Overlapping Slicing

Rani V. Ingawale, Ms. D. A. Chaudhari, DYPCOE - Akurdi, Savitribai Phule Pune University

***Abstract*—**Data Mining is the process of analysing data from different perspective, summarizing it into useful information and extracts the needed information from the database. Most enterprises are collecting and storing data in large database. Database privacy is a important responsibility of organizations for to protects clients sensitive information, because their clients trust them to do so. Various anonymization techniques have been proposed for the privacy of sensitive microdata. However, there is considered between the level of privacy and the usefulness of the published data. Recently, slicing was proposed as a technique for anonymized published dataset by partitioning the dataset vertically and horizontally. This paper proposes a technique to increase the utility and privacy of a sliced dataset by allowing overlapped slicing while maintaining the prevention of membership disclosure. Also provide secure data access for multiple domains. This novel approaches work on overlapped slicing to improve preserve data utility and privacy better than traditional slicing.

Index Terms*—Data anonymization, Privacy preservation, Data publishing, Data security*

*****

INTRODUCTION

Many organizations collecting information is used for knowledge-based decision making and analysis purpose. So, there is need to share the information. But, the data holds some sensitive information and people does not want their sensitive information to be revealed. Sharing data in its original form thus reveal the individual privacy. So, to prevent this violation of privacy there should be some technique to publish the data in such a way that privacy is preserved and at the same time data analysis can be done effectively. Microdata holds information about an individual entity, like a person, or-ganization. Several microdata anonymization techniques have been proposed to protect sensitive attributes. The most popular one is generalization and bucketization. Here, data attributes are partitioned into three 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 at-tributes 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 break 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. Overlapped slicing increase the utility and Privacy of a sliced dataset with multiple sensitive attributes in different domains.

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.

C. Aggarwal[2] initially proposed on k-anonymity and curse of dimensionality concept. The author proposed privacy pre-serving 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], [14] 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 [15]. 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.

III. IMPLEMENTATION DETAILS

A. Slicing:

Slicing divides the data set both horizontally and vertically. 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. In horizontal partition grouping tuples into buckets. Within bucket, column values are randomly permuted to break the linking between different columns. 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 preserves utility, and preserves the correlations among such attributes. Slicing preserve 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 tuple. There are generally various matching 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 and horizontally. Data in which have multiple sensitive attributes used slicing for membership revelation protection and conserves good data useful than generalization and bucketization.

TABLE I: The Original Table

Age	GenderZipcode	Occupation	Education	Disease	Political Opine-Ion	Crime
20	F 12578	Student	12th	Flu	BGP	Robbery
41	M 12589	Government	PG	Cancer	Aap	Null
26	M 12460	Sales	10th	Cancer	Congress	Thief
23	F 12216	Student	Graduate	Flu	BGP	Null
29	M 12903	Agriculture	12th	diabetes	Congress	Thief
32	M 12093	Army	Graduate	hypertension	Shivsena	Null

B. Privacy Threats

When publishing microdata, there are three types of information disclosure threats.

1) Membership Disclosure Protection- The first type is membership disclosure, when the data to be published is selected from a larger 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. [7]
2) Identity Disclosure Protection- The second type is identity disclosure, which occurs when an individual is linked to a particular record in the released table. In some situations, one wants to protect against identity disclosure when the attackers is undefined of membership.[9]
3) 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

 ISSN: 2321-8169
1992 - 1997

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.[9]

TABLE II: The Sliced table

(Gender,Occupation)	(Zipcode,Education)	(Age, Disease)
(M,Sales)	(12460,10th)	(32,hypertention)
(M,Army)	(12578,12th)	(26, Cancer)
(F,Student)	(12093,Graduate)	(20,Flu)
(M,Agriculture)	(12216,Graduate)	(29,diabetes)
(F,Student)	(12589,PG)	(23, Flu)
(M,Government)	(12903,12th)	(41,Cancer)

C. Improved Slicing

Here, a novel data anonymization model is lead that improves limitations of slicing. The major influences of this model are the use of an overlapped clustering technique [16] in the attribute partitioning phase and the use of an alternative tuple partitioning algorithm. Improved slicing works by first finding the correlations between each pair of attributes and then clustering these attributes into columns by overlapped clustering on the basis of their association coefficients. The dataset is then horizontally partitioned into buckets satisfying l-diversity [8] using a novel tuple partitioning algorithm. The columns within each bucket are then arbitrarily permuted with respect to one another to give an enhanced sliced dataset.

D. Overlapped Slicing

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. 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 III show the overlapped slicing technique. In Table II, Age is grouped with Disease and Occupation is grouped with Gender. Even if Occupation also had a nearly high correlation with Disease but Gender did not, they could not be joined into a higher group and thus the data utility because the association between Disease and Occupation is gone. In Table III, the attributes Occupation and Disease are existing in more than one column means they are overlapping. This allows highly associated attributes to group together. This also solves the problem of singular columns by merging associated attributes into a different column instead of just leaving out an attribute with a low correlation. The idea of Overlapping Correlation Clustering[16] was suggested by F. Bonchi et al. and can be occupied to the attribute partitioning phase of the slicing algorithm.

In this technique, a set of non-overlapping clusters is renewed to overlapped clusters by permitting an attribute to belong to more than one cluster by examining the equivalent function between the attribute and each cluster.

TABLE III: The Overlapped Sliced Table

(Gender,Occupation)	(Zip,Education)	(Age, Disease)	(Disease,Occupation)
(M,Sales)	(12460,10th)	(32,Hypertention)	(Dyspepsis,Sales)
(M,Army)	(12578,12th)	(26, Cancer)	(Flu,Student)
(F,Student)	(12093,Graduate)	(20,Flu)	(Hypertention, Army)
(M,Agriculture)	(12216,Graduate)	(29,diabetes)	(Diabetes,Agriculter)
(F,Student)	(12589,PG)	(23, Flu)	(Cancer,Government)
(M,Government)	(12903,12th)	(41,Cancer)	(Flu,Student)

-{s}].

The probability that t is in bucket B is =
$p(t, B) = \frac{f(t; B)}{}$; f(t) =f(t, B); p = (s|t, B) = D(t, B).[s]
(2) Where D(t, B) = Distribution of the candidate sensitive values in B.
D(t,B).[s] = The probability sensitive value s in the distribution.

F. System Overview

Architecture of the proposed system is shown in figure 1.

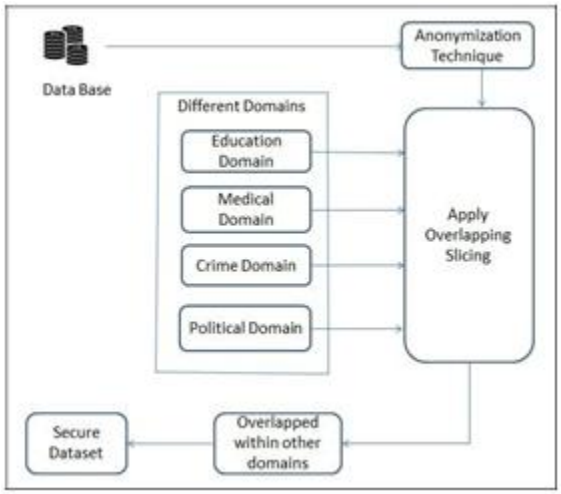

Fig. 1: System Architecture

 ISSN: 2321-8169
1992 - 1997

E. Mathematical Model

1) T = Microdata 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

$$p(t; s) = p(t; B)p(sjt; B) \quad (1)$$

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 = $t[C_1]$, $t[C_2]$….$[C_c]$
B's column value = $B[C_1]$, $B[C_2]$….$B[C_c]$
$f_i(t, B)$ = Fraction of occurrences of $t(C_i)$ in $B(C_i)$:

$f_c(t,B)$ = Fraction of occurrences of $t[C_c - \{s\}]$ in $B[C_c$

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.

G. Experimental Setup

a) Software Requirement : Basic software specifications are:

1 Interface

– Hard disk: 40 GB

– RAM: 512 MB

– Processor Speed: 3.00GHz

– Processor: Pentium IV Processor

2 Software Interface

– Language used - Java (Java Development Kit JDK 1.5)

– Operating System - Windows XP/2007

IV. RESULTS AND DISCUSSION

Overlap slicing with multiple sensitive 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 be-tween 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.

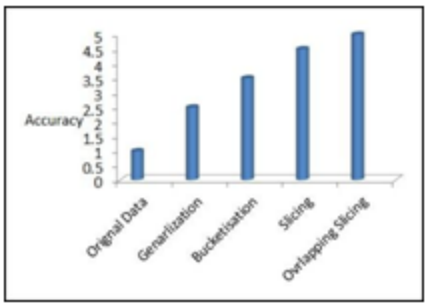

Fig. 2: Graph For Privacy

In Table IV show the Overlapped slicing with Education Domain. Here full access to education domain and over-lapped with other domains.

 ISSN: 2321-8169
1992 - 1997

TABLE IV: The Overlapped Slicing Table with Eduction Domain

Age	Gender	Zipcode	Occupa-Tion	Eduction	(Political-Opinion, Disease)	(Disease,Crime)
20	F	12578	Student	12th	(Congress,Cncer)	(Cancer,theaf)
41	M	12589	Govern.	PG	(BGP,Flu)	(Flu,Robbery)
26	M	12460	Sales	10th	(Aap,Cancer)	(Cancer,Null)
23	F	12216	Student	Graduate	(Shivsena,diabetes)	(hyperT,Null)
29	M	12903	Agri.	12th	(Congress,hyperT)	(Flu,Null)
32	M	12093	Army	Graduate	(BGP,Flu)	(diabetes,Theaf)

In Table V show the Overlapped slicing for Medical do-main. Here full access to medical domain and overlapped with other domains.

TABLE V: The Overlapped Slicing table with Medical Domain

Age	Gen-Der	Zip-code	Disease	(PoliOpin, Eduction)	(Occupation, Crime)	(Occupation, Eduction)	(PoliOpin, Crime)
20	F	12578	Flu	(Congress,10th)	(Sale,theaf)	(Gornm,PG)	(Aap,Null)
41	M	12589	Cancer	(BGP,12th)	(Student,Robbery)	(Sale,10th)	(congress,theaf)
26	M	12460	Cancer	(Aap,PG)	(Gov,Null)	(Student,12th)	(BGP,Robbery)
23	F	12216	Flu	(Shivsena,12th)	(Army,Null)	(Agri,12th)	(Shivsena,Theaf
29	M	12903	Dibetes	(Congress,Graduate)	(Student,Null)	(Army,Graduate)	(Congress,Null)
32	M	12093	HyerT	(BGP,Graduate)	(Agri,theaf)	(Student,Graduate)	(BGP,Null)

In Table VI show the Overlapped slicing for Crime domain. Here full access to Crime domain and overlapped with other domains.

In Table VII show the Overlapped slicing for Political domain. Here full access to Political domain and overlapped with other domains.

TABLE VI: The Overlapped Slicing Table with Crime Domain

Age	Gen-der	Zip-code	Crime	(PoliOpin, Eduction)	(Occupation, Disease)	(Occupation, Education)	(PoliOpin, Disease)
20	F	12578	Robbery	(Congress,10th)	(Sale,Cancer)	(Gornm,PG)	(Aap,Cancer)
41	M	12589	Null	(BGP,12th)	(Student,Flu)	(Sale,10th)	(Congress,Cancer)
26	M	12460	Theaf	(Aap,PG)	(Gove,Cancer)	(Student,12th)	(BGP,Flu)
23	F	12216	Null	(congress,Graduate)	(Army,HyperT)	(Agri,12th)	(Shivsena,Diabetes
29	M	12903	Theaf	(Shivsena,12th)	(Student,Flu)	(Army,Graduate)	(Congress,HypeT)
32	M	12093	Null	(BGP,Graduate)	(Agri,Diabetes)	(Student,Graduate)	(BGP,Flu)

TABLE VII: The Overlapped Slicing Table with Political-Opinion Domain

Age	Gen-der	Zip-code	Political - Opinion	(Crime, Educ-tion)	(Occupation, Disease)	(Occupation, Education)	(Crime, Dis-ease)
20	F	12578	BGP	(Theaf,10th)	(Sale, Cancer)	(Gornm,PG)	(Cancer,theaf)
41	M	12589	Aap	(Robbery,12th)	(Student,Flu)	(Sale,10th)	(Flu,Robbery)
26	M	12460	Congress	(Null,PG)	(Gove,Cancer)	(Student,12th)	(Cancer,Null)
23	F	12216	BGP	(Null,Graduate)	(Army,HyperT)	(Agri,12th)	(Hyper,Null)
29	M	12903	Shivsena	(Theaf,12th)	(Student,Flu)	(Army,Graduate)	(Flu,Null)
32	M	12093	congress	(Null,Graduate)	(Agri,Diabetes)	(Student,Graduate)	(Diabetes,theaf)

V. CONCLUSION

This paper presents a novel technique for increasing the utility of anonymized datasets by improving of slicing. Overlapped slicing can duplicate an attribute in more than one column and this leads to greater data utility because of an increased release of attribute correlations. Overlapped slicing satisfies all the privacy safeguards of traditional slicing such as prevention of attribute disclosure and membership disclosure. Here overlapped slicing demonstrate the greater data utility provided by improved slicing while satisfying l-diversity.

ACKNOWLEDGMENT

I take this opportunity to express my profound gratitude and deep regards to my guide Ms. D. A. Chaudhari 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 Mrs. M. A. Potey, PG Coordinator Mrs. S. S. Pawar, 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, Member, 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 privdestruction 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 5

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

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

[8] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu, Privacy Preserving Data Publishing Concepts and Techniques ,Data mining and knowledge discovery series 2010.

[9] 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

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

[11] 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.

[12] 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.

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

[14] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang,Aggregate Query Answering on Anonymized Tables, Proc. IEEE 23rd Int?l Conf. Data Eng. (ICDE), pp1. 116-125, 2007.

[15] S.Kiruthika, Dr.M.Mohamed Raseen, Enhanced Slicing Models for Preserving Privacy in Data Publication, ICCETET, pages 406-409 in proc. of IEEE, 2013.

[16] F.Bonchi, A. Gionis,and A. Ukkonen, "Overlapping Correlation Clustering" in proceeding IEEE 11th International conference on Data Mining, 2011,PP.51-60

[17] Lan Sun, Yilei Wang, Yingjie Wu,A Survey of Transaction dada Anonymous publication in IEEE Symposium on Robotics and Applications 2012.

Rani V. Ingawale received the B.E. degree in Computer Science from SVERI College of Engineering Pandharpur in 2009. Now she is pursuing Master degree in Computer Engineering from D.Y.Patil College of Engineering Akurdi, Pune.

Dipalee A. Chaudhari received the BE degree in Computer Science and Engineering from University of Pune in 2000 and ME in Computer Engineering from University of Pune in 2010 and has 8 years of teaching experience. She is currently working as Assistant professor at D. Y. Patil College of Engineering, Akurdi, Pune.