The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 10||Pages|| 01-04||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

Enhanced Slicing For Privacy Preserving Data Publishing

[1,]K.Vani , [2,] B.Srinivas
[1,]*Department of CSE KITS Warangal, AP.*
[2,] *Assistant Professor Department of CSE KITS,Warangal*

----------ABSTRACT----------
The anonymization tactics, such as generalization along with bucketization, are actually designed regarding privacy conserving microdata creating. Recent work shows that generalization loses quite a bit of information, especially regarding high dimensional data. Bucketization, on the other hand, does not really prevent member's program disclosure and does not apply regarding data that will not have a distinct separation between quasi-identifying characteristics and hypersensitive attributes. In this paper, all of us present any novel technique called Slicing, which partitions the information both flat in a trench and vertically. We show that slicing preserves superior data power than generalization and can be used for member's program disclosure security. Another important good thing about slicing is that it can manage high-dimensional data. An extension may be the notion of overlapping chopping, which duplicates a feature in more than one columns. This particular releases a lot more attribute correlations. Here we provide the overview of the vertical partitioning of the dataset ,Highly correlated attributes are grouped as columns that will not reveals the individuals information. frequent occurrence of the attribute values has no identification risks. To do this we have to used the one of the data mining algorithm "Apriori " .

KEYWORDS*: Publishing, Privacy, Data Security*

Date of Submission: 1, October, 2013 Date of Acceptance: 20, October 2013

I. INTRODUCTION

Privacy preserving data mining is a novel research in data mining. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the individuals information is not revealed as well as the data must be useful for analyzing characteristics of population. For example the Data publisher has a table of attributes like *(Explicit Identifier, Quasi Identifier, Sensitive Attributes,*Non-Sensitive Attributes).The main consideration in privacy preserving data mining is two fold. First, explicit identifiers Name, address and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. second sensitive information which can be mined from a database by using data mining algorithms. Even with all explicit identifiers being removed also there is a privacy threat linking to reidentifying the record owner. To control the membership disclosure in publishing of microdata needs privacy preserving techniques such as generalization for k-anonymity and bucketization for l-diversity. In this paper we have used a creative privacy technique called slicing .slicing prevents membership disclosure and attribute disclosure. slicing partitions the data set both vertically and horizontally.rest of the paper is organized as follow :in section ii the related work of generalization and bucketization is given. We define slicing and the proposed research work in section iii. We formalize the results in section iv and conclude the paper in section v.

II . RELATED WORK

Two popular anonymization techniques are generalization and bucketization . Generalization replaces low level values with high level concept values and transforms the QI-values in to Range of values. so that the tuples in the same bucket can't be distinguish by their QI values. If one record in the table has some value QI atleast (k-1) other records also have the value QI each record is indistinguishable from atleast (k-1) other records with respect to QI. K-anonymity provides protection against membership disclouser. But the generalization for k-anonymity losses information especially for high dimensional data. K-anonymity not handles high dimensional data and it can't prevent attribute disclosure it reduces the data utility of the generalized data.Each attribute is generalized separately, the correlation between different attributes are lost.Bucketization partitions the tuples intobuckets.but it has several limitations .

First bucketization does not prevent membership disclosure.because bucketization publishes the QI values in their original forms ,an adversary can easily finds the individuals information from the published data. Second bucketization requires a clear separation between QIs and SAsA novel data anonymization technique called slicing overcomes the above limitations of generalization and bucketization .In this paper we have done the slicing with vertical partitioning of the dataset and Horizontal partitioning of the dataset .

II. PROPOSED WORK

slicing partitions the dataset both horizontally and vertically. Vertical partitioning means grouping the attributes into columns based on the correlations of the attributes .Attribute partitioning enables slicing handles the high dimentional data.Horizontal partitioning is done by grouping the tuples into buckets and within each bucket, values in each column are randomly permuted to break the linking between different columns. Slicing breaks the association cross columns, but preserves the associations within each column. Slicing provide privacy protection ,that the slicing process ensures that for any tuple there are multiple matching buckets. Given a tuple t, c is the number of columns and v is the values of the column, a bucket is a matching bucket for t if and only if for each column value appears atleast once in the I th column of the bucket. Any bucket that contains the original tuple is the matching bucket .At the same the matching bucket contains other tuples with same values but not all. Slicing prevents membership disclosure by dividing the QI attributes into different columns and correlations among different columns are broken. Each columns attribute values are randomly permuted. No one can tell the tuple is in the original table. Slicing can be used to prevent attribute disclosure based on the privacy requirement of l-diverse slicing . if a bucket is l-diverse then there are l or more well represented values for SAs.
Let T be the dataset table which is sliced by applying slicing algorithm .This algorithm has two phases: vertical partitioning of the dataset and horizontal partitioning of the dataset.

VERTICAL PARTITIONING OF THE DATA

Vertical partitioning of the dataset consists of subset of attributes made as columns. Table T contains d attributes A={A1,A2,…..Ad} and c columns c1,c2,…cn. Our algorithm grouping the highly correlated attributes are in the same column . grouping the highly correlated attributes preserves the data utility that means data should maintain privacy and data is useful for analysis .the association of uncorrelated attributes provide more identification risk. the

highly correlated attribute values are more frequent than the uncorrelated attributes and thus less identification risks. For finding frequent values of the attributes we have used the closed Apriori enhancement algorithm .

i)Closed Apriori Enhancement: In this enhancement we consider domain of each attribute as the itemset and implements the enhancement as an Apriori like algorithm that directly mines frequent closed itemsets. There are two main steps in Close. The first is to use bottom-up search to identify generators, the smallest frequent itemset that determines a closed itemset. All generators are found using a simple modification of Apriori. After finding the frequent sets at level k, Close compares the support of each set with its subsets at the previous level. If the support of an itemset matches the support of any of its subsets, the itemset cannot be a generator and is thus pruned. The second step in Close is to compute the closure of all the generators found in the first step. To compute the closure of an itemset we have to perform an intersection of all transactions where it occurs as a subset. The closures for all generators can be computed in just one database scan, provided all generators fit in memory. Nevertheless computing closures this way is an expensive operation.

ii) Vertical Overlapping Slicing

In slicing where each attribute is exactly in one column. In this approach we consider the sensitive attribute to combines this attribute at different sets to provide enhanced anonymity .By duplicating the attribute in more than one column provides better data utility but this releases more attribute correlations. But the privacy implications must be carefully maintained.

HORIZONTAL PARTITIONING OF THE DATASET

Horizontal partitioning of the dataset is nothing but the partitioning of the tuples into buckets. subset of tuples is called as buckets tuple partitioning algorithm is first divides the tuples into buckets and to check whether the each bucket satisfies the l-diversity or not. For bucketization slicing uses automatic bucketization algorithm is one of the sub module in the tuple partitioning algorithm. Slicing prevents attribute disclosure by using the privacy requirement l-diversity. Diversity-check algorithm is used for l-diversity.

i)Automatic Bucketization

In this approach we consider the average of min age and the max age at each stage of bucketization. the process continues until the left or the right side of each split bucket is less than equal to some threshold value of the number of tuples. This approach work the same way as the binary search algorithm but splits the tuples in to different buckets.

ii)l-diversity slicing

The main part of the tuple partitioning algorithm is to check whether the each tuples of the sliced table satisfies the l-diversity.for each tuple t the algorithm maintains a list of probabilities of the tuples matching bucket p(t,B) and the distribution of candidate sensitive values D(t,B). The algorithm first takes one scan of each bucket B to find the matching of tuples column values with the buckets column values and record their matching probability p(t,B) and then compute the probability that tuple t takes sensitive value s within the tuples matching bucket B i.e D(t,B). A final scan of the tuples in the dataset will compute the p(t,s) values based on the law of total probability.

$P(t,s) = \sum p(t,B) * D(t,B)[s]$.

The sliced table is l-diverse iff for all value s, $p(t,s) \leq 1 / l$. if $p(t,s) \geq 1/l$ then the adversary cannot correctly learn the sensitive value of any individual with probability greater than 1/l.

IV. RESULTS

The concept of this paper is implemented and different results are shown below, The proposed paper is implemented in Java technology on a Pentium-IV PC with minimum 20 GB hard-disk and 1GB RAM. The propose paper's concepts shows efficient results and has been efficiently tested on different Datasets.

Age	WorkClass	FinalWeight	Education	Edu-Num	Marital Sta...	Occupation	Relations...	Race	Sex	Capital Gain	Capital Lo...	Hours per ...	Country	Salary	Title 16
[17,28]	P	226802	11	7	NM	MOI	Own-child	BL	F	0	0	40	United-St...	<=50K.	
[17,28]	LG	336951	AA	12	MCS	PRS	Husband	WH	F	0	0	40	United-St...	>50K.	
[17,28]		103497	SC	10	NM		Own-child	WH	F	0	0	30	United-St...	<=50K.	
[17,28]	P	369667	SC	10	NM	OS	Unmarried	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	82091	HS	9	NM	AC	Not-in-fa...	WH	F	0	0	39	United-St...	<=50K.	
[17,28]	SG	444554	SC	10	NM	OS	Own-child	WH	F	0	0	25	United-St...	<=50K.	
[17,28]	P	220931	B	13	NM	PS	Not-in-fa...	WH	F	0	0	43	Peru	<=50K.	
[17,28]	P	205947	B	13	MCS	PS	Husband	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	236427	HS	9	NM	AC	Own-child	WH	F	0	0	20	United-St...	<=50K.	
[17,28]	P	134446	HS	9	SP	MOI	Unmarried	BL	F	0	0	54	United-St...	<=50K.	
[17,28]	SENI	188274	B	13	NM	SA	Not-in-fa...	WH	F	0	0	50	United-St...	<=50K.	
[17,28]	LG	258120	SC	10	MCS	PRS	Husband	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	43311	HS	9	DV	EM	Unmarried	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	248446	56	3	NM	PHS	Not-in-fa...	WH	F	0	0	50	Guatemala	<=50K.	
[17,28]	P	269430	10	6	NM	MOI	Not-in-fa...	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	257509	HS	9	NM	CR	Own-child	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	SG	138371	SC	10	NM	FF	Own-child	WH	F	0	0	32	United-St...	<=50K.	
[17,28]	P	242832	AV	11	MCS	PS	Wife	WH	F	0	0	36	United-St...	>50K.	
[17,28]	P	54440	SC	10	NM	OS	Own-child	WH	F	0	0	20	United-St...	<=50K.	
[17,28]	P	214399	SC	10	NM	OS	Own-child	WH	F	0	1721	24	United-St...	<=50K.	
[17,28]	P	54164	HS	9	NM	OS	Not-in-fa...	WH	F	14084	0	60	United-St...	>50K.	
[17,28]	P	110677	SC	10	NM	AC	Own-child	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	31208	MS	14	NM	EM	Not-in-fa...	WH	F	0	0	40	United-St...	<=50K.	
[17,28]	P	105460	SC	10	NM	OS	Own-child	WH	F	0	0	20	United-St...	<=50K.	
[17,28]	P	388946	SC	10	SP	HC	Not-in-fa...	WH	F	0	0	40	United-St...	<=50K.	

Fig. 1 Proposed system performing Anonymity on the dataset.

Age	Workclass	FinalWeight	Education	Edu-Num	Marital Sta...	Occupation	Relations...	Race	Sex	Capital Gain	Capital Lo...	Hours per ...	Country		Salary
25	P	226802	11	7	NM	MOI	Own-child	BL	F	0	0	40	United-St...		<=50K.
28	LG	336951	AA	12	MCS	PRS	Husband	WH	F	0	0	40	United-St...		<=50K.
18		103497	SC	10	NM		Own-child	WH	F	0	0	30	United-St...		<=50K.
24	P	369667	SC	10	NM	OS	Unmarried	WH	F	0	0	40	United-St...		>50K.
26	P	82091	HS	9	NM	AC	Not-in-fa...	WH	F	0	0	39	United-St...		<=50K.
20	SG	444554	SC	10	NM	OS	Own-child	WH	F	0	0	25	United-St...		<=50K.
25	P	220931	B	13	NM	PS	Not-in-fa...	WH	F	0	0	43	Peru		<=50K.
25	P	205947	B	13	MCS	PS	Husband	WH	F	0	0	40	United-St...		<=50K.
22	P	236427	HS	9	NM	AC	Own-child	WH	F	0	0	20	United-St...		<=50K.
23	P	134446	HS	9	SP	MOI	Unmarried	BL	F	0	0	54	United-St...		<=50K.
24	SENI	188274	B	13	NM	SA	Not-in-fa...	WH	F	0	0	50	United-St...		<=50K.
23	LG	258120	SC	10	MCS	PRS	Husband	WH	F	0	0	40	United-St...		<=50K.
26	P	43311	HS	9	DV	EM	Unmarried	WH	F	0	0	40	United-St...		<=50K.
22	P	248446	56	3	NM	PHS	Not-in-fa...	WH	F	0	0	50	Guatemala		<=50K.
17	P	269430	10	6	NM	MOI	Not-in-fa...	WH	F	0	0	40	United-St...		<=50K.
20	P	257509	HS	9	NM	CR	Own-child	WH	F	0	0	40	United-St...		<=50K.
20	SG	138371	SC	10	NM	FF	Own-child	WH	F	0	0	32	United-St...		<=50K.
28	P	242832	AV	11	MCS	PS	Wife	WH	F	0	0	36	United-St...		<=50K.
18	P	54440	SC	10	NM	OS	Own-child	WH	F	0	0	20	United-St...		<=50K.
21	P	214399	SC	10	NM	OS	Own-child	WH	F	0	1721	24	United-St...		<=50K.
22	P	54164	HS	9	NM	OS	Not-in-fa...	WH	F	14084	0	60	United-St...		<=50K.
21	P	110677	SC	10	NM	AC	Own-child	WH	F	0	0	40	United-St...		<=50K.
26	P	31208	MS	14	NM	EM	Not-in-fa...	WH	F	0	0	40	United-St...		<=50K.
19	P	105460	SC	10	NM	OS	Own-child	WH	F	0	0	20	United-St...		<=50K.
21	P	388946	SC	10	SP	HC	Not-in-fa...	WH	F	0	0	40	United-St...		>50K.

Fig. 1 Proposed system performing Diversity on different dataset

Set 1	Set 2	Set 3	Set 4	Set 5
(25,P, 226802)	(11, 7,NM)	(MOI, Own-child,BL)	(F , 0, 0,)	(40, United-States, <=50K.)
(28,LG, 336951)	(AA, 12,MCS)	(PRS, Husband,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(18,, 103497)	(SC, 10,NM)	(, Own-child,WH)	(F , 0, 0,)	(30, United-States, <=50K.)
(24,P, 369667)	(SC, 10,NM)	(OS, Unmarried,WH)	(F , 0, 0,)	(40, United-States, >50K.)
(26,P, 82091)	(HS, 9,NM)	(AC, Not-in-family,WH)	(F , 0, 0,)	(39, United-States, <=50K.)
(20,SG, 444554)	(SC, 10,NM)	(OS, Own-child,WH)	(F , 0, 0,)	(25, United-States, <=50K.)
(25,P, 220931)	(B, 13,NM)	(PS, Not-in-family,WH)	(F , 0, 0,)	(43, Peru, <=50K.)
(25,P, 205947)	(B, 13,MCS)	(PS, Husband,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(22,P, 236427)	(HS, 9,NM)	(AC, Own-child,WH)	(F , 0, 0,)	(20, United-States, <=50K.)
(23,P, 134446)	(HS, 9,SP)	(MOI, Unmarried,BL)	(F , 0, 0,)	(54, United-States, <=50K.)
(24,SENI, 188274)	(B, 13,NM)	(SA, Not-in-family,WH)	(F , 0, 0,)	(50, United-States, <=50K.)
(23,LG, 258120)	(SC, 10,MCS)	(PRS, Husband,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(26,P, 43311)	(HS, 9,DV)	(EM, Unmarried,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(22,P, 248446)	(56, 3,NM)	(PHS, Not-in-family,WH)	(F , 0, 0,)	(50, Guatemala, <=50K.)
(17,P, 269430)	(10, 6,NM)	(MOI, Not-in-family,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(20,P, 257509)	(HS, 9,NM)	(CR, Own-child,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(20,SG, 138371)	(SC, 10,NM)	(FF, Own-child,WH)	(F , 0, 0,)	(32, United-States, <=50K.)
(28,P, 242832)	(AV, 11,MCS)	(PS, Wife,WH)	(F , 0, 0,)	(36, United-States, <=50K.)
(18,P, 54440)	(SC, 10,NM)	(OS, Own-child,WH)	(F , 0, 0,)	(20, United-States, <=50K.)
(21,P, 214399)	(SC, 10,NM)	(OS, Own-child,WH)	(F , 0, 1721,)	(24, United-States, <=50K.)
(22,P, 54164)	(HS, 9,NM)	(OS, Not-in-family,WH)	(F , 14084, 0,)	(60, United-States, <=50K.)
(21,P, 110677)	(SC, 10,NM)	(AC, Own-child,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(26,P, 31208)	(MS, 14,NM)	(EM, Not-in-family,WH)	(F , 0, 0,)	(40, United-States, <=50K.)
(19,P, 105460)	(SC, 10,NM)	(OS, Own-child,WH)	(F , 0, 0,)	(20, United-States, <=50K.)
(21,P, 388946)	(SC, 10,SP)	(HC, Not-in-family,WH)	(F , 0, 0,)	(40, United-States, >50K.)

Fig. 3 Proposed system performing Slicing on different datasets

V. Conclusions

This paper presents a new approach called slicing to privacy preserving micro data publishing. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. We illustrate how to use slicing to prevent attribute disclosure and membership disclosure. Our experiments show that slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. The general methodology proposed by this work is that: before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization. The rationale is that one can design better data anonymization techniques when we know the data better.

REFERENCES

[1] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.

[2] R.C.-W. Wong, A.W.-C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 543-554, 2007.

[3] T. Li, N. Li, and J. Zhang, "Modeling and Integrating Background Knowledge in Data Anonymization," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 6-17, 2009.

[4] T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 446-455, 2008.

[5] C. Dwork, "Differential Privacy: A Survey of Results," Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.

[6] 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.

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

[8] A. Inan, M. Kantarcioglu, and E. Bertino, "Using Anonymized Data for Classification," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[9] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.

[10] M.E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 665-676, 2007.

[11] T. Li and N. Li, "On the Tradeoff between Privacy and Utility in Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.

[12] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-Preserving Anonymization of Set-Valued Data," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008.

[13] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[14] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.

[15] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, "Anonymizing Transaction Databases for Publication," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.