ISSN:0975-9646

Efficient Technique for Annonymized Microdata Preservation using Slicing

Manjusha S. Mirashe[1] Prof. Kapil N. Hande[2]

[1.] M. Tech. CSE Department, Priydarshani Bhagwati Chaturvedi College of Engineering, Nagpur.
[2.] Asst. Prof. CSE Department, Priydarshani Bhagwati Chaturvedi College of Engineering, Nagpur.

Abstract: **Data Anonymization is always being a subject of researchers in the last few years. Privacy Preserving Data Mining, i.e. the study of data mining side-effects on privacy, which receives an increasing attention from the research community. Privacy-preservation data publishing has received lot of attention, as it is always a problem of how to secure a database of high dimension. In much organization where large number of confidential data is available, such data must be secured. The personal data may be misused, for a variety of purposes. In order to alleviate these concerns, a number of techniques have recently been proposed in order to perform the data mining tasks in a privacy-preserving way. There are several anonymization techniques available such as generalization and bucketization that are designed for privacy preservation of microdata publishing. But it has been seen that for high dimension data generalization looses the information, bucketization on other hand does not prevent membership disclosure. We present another anonymization technique known as Slicing. The significance of using slicing is that it can handle high dimension data. Slicing preserves better data utility than generalization and also prevents membership disclosure. This paper focus on effective method that can be used for providing better data utility and can handle high-dimensional data.**

Keywords- Data anonymization, Data publishing, Data security, Privacy Preservation, Privacy Threats

I. INTRODUCTION

Today most of the organizations need to publish microdata. Microdata contain records each of which contains information about an individual entity, such as a person or a household. Many microdata anonymization techniques have been proposed and the most popular ones are generalization with k-anonymity and bucketization with l-diversity. In both methods attributes are into three categories, some of them are identifiers that can be uniquely identified such as Name or security number, some are quasi-identifiers. These quasi–identifiers are set of attributes are those that in combination can be linked with the external information to reidentify such as birthdate, sex and zip code and the third category is sensitive attributes, this kind of attributes are unknown to the opponent and are considered sensitive such as disease and salary. These are three categories of attributes in microdata. In both the anonymization techniques first identifiers are removed from the data and then partitions the tuples into buckets. Generalization transforms the quasi-identifying values in each bucket into less specific and semantically constant so sthat tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SA values from the QI values by randomly permuting the SA values in the bucket .The anonymized data consist of a set of buckets with permuted sensitive attribute values. The identity of patients must be protected when patient data is shared. Previously we used techniques using k-anonymity and l-diversity. Existing works mainly considers datasets with a single sensitive attribute while patient data consists multiple sensitive attributes such as diagnosis and treatment. So both techniques are not so efficient for preserving patient data. So, we are presenting a new technique for preserving patient data and publishing by slicing the data both horizontally and vertically. Data slicing can also be used to prevent membership disclosure and is efficient for high dimensional data and preserves better data utility.

II. RELATED WORK

To improve the disclosure of the patient data and to preserve better data utility sliced data is more efficient when compared to generalization and bucketization. In case of generalization [9] , it is shown that generalization loses considerable amount of information especially for high dimensional data. In order to perform data analysis or data mining tasks on the generalization table, the data analyst has to make the uniform distribution assumption that every value in a generalized set is equally possible and no other distribution assumption can be justified[11]. This significally reduces the data utility of the generalized data. In generalizes table each attribute is generalized separately, correlations between different attributes are lost. This is an inherent problem of generalization. In case of bucketization, it has better data utility than generalization but does not prevent membership disclosure. Secondly bucketization publishes the QI values in their original forms, an opponent can easily find out whether an individual has a record in the published data or not. This means that membership information of most individuals can be inferred from the bucketization table. Also bucketization requires clear separation between QI and SI values .By separating the sensitive attributes from the quasi-identifying attributes, bucketization breaks the attribute correlation between the QIs and SAs. However in many data sets it is unclear that which attributes are QI's and which are SA's.

So, bucketization is also not so efficient for preserving microdata and publishing. Slicing has some connections to marginal publication [15], both of them release correlations among a set of attributes. Slicing is quite different from marginal publication. First, marginal publication can be viewed as a special case of data slicing which does not have horizontal partitioning. Therefore correlations among attributes in different columns are lost in marginal publication.

III. Basic Idea of Data Slicing

In this paper, we introduce a new method, called DATA SLICING. This method partitions the data both horizontally and vertically. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. At last, within each bucket, values in each column are randomly permutated to break the association between different columns. The core idea of data slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better data utility than bucketization and generalization. Data analysis methods such as query answering can be easily viewed on sliced data. Data slicing method consists of four stages. They are

1. Partitioning attributes and columns
2. Partitioning tuples and buckets.
3. Generalization of buckets
4. Matching the buckets.

In the first stage, an attribute partition consists of several subsets of A, where each attribute belongs to exactly one subset. A column is nothing but a subset of attributes. Consider only one sensitive attribute S, if the data contains multiple sensitive attributes, one can either consider them separately or consider their joint distribution [13]. The column that contain sensitive attribute is called as the sensitive column. Remaining column contains only quasi identifying attributes. In the second stage, partitioning of tuples is taken place, each tuple belongs to exactly one subset and the subset of tuples is called a bucket. In the third stage, column generalization is done. A column generalization maps each value to the region in which the value is contained. In the last stage we have to check whether the buckets are matching.

IV. Proposed Work

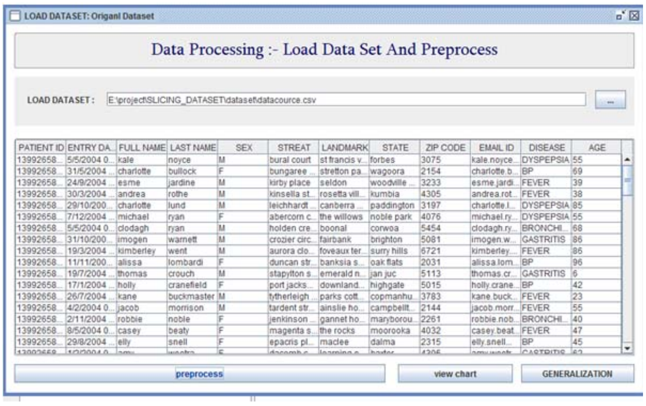
1. Problem Statement: Database privacy is a concept that is important to organizations and private citizens alike. Privacy professionals also can secure storage systems against theft involving servers, hard drives, desktops and laptops. Organizations should ensure that storage management interfaces and all database backups, whether on-site or off-site, maintain their integrity. If attacks on a database occur, it is an organization's responsibility to take defensive measures. This might first entail the immediate classification of data according to importance. Then, encryption methods might be employed to help protect applications and data based on their sensitivity levels. Of course, the best method of protecting a database's privacy is prevention. One method of database privacy protection might include assessing a database regularly for exploits and signs that it has been compromised. If an organization can detect exploits or indications of database compromising before the threat becomes real and unmanageable, the database might be able to be rectified with little and reversible damage.

2. Goals: An important research problem is for handling high-dimensional data. As per the above, Privacy Preservation for high dimensional database is important. There are two popular data anonymization technique Generalization and Bucketization. These techniques are designed for privacy preserving microdata publishing. Our Proposed work includes a slicing technique which is better than generalization and bucketization for the high dimension data sets.Slicing preserves better data utility than generalization and can be used for membership disclosure protection.

V. Data Slicing

1. Overview

The overall method of slicing has been discussed above. The original microdata consist of quasi identifying values and sensitive attributes. In figure 1 patient data in a hospital. The data consists of Age, Address, Id, Sex, Zipcode, disease. Here the QI values are {age, sex, zipcode} and the sensitive attribute is {disease}.A generalized table replaces values.



Data Processing :- Load Data Set And Preprocess

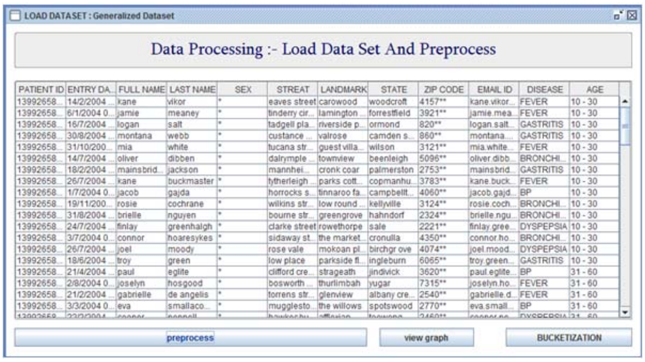
LOAD DATASET : E:\project\SLICING_DATASET\dataset\datacource.csv

PATIENT ID	ENTRY DA...	FULL NAME	LAST NAME	SEX	STREAT	LANDMARK	STATE	ZIP CODE	EMAIL ID	DISEASE	AGE
13992658...	5/5/2004 0...	kale	noyce	M	bural court	st francis v...	forbes	3075	kale.noyce...	DYSPEPSIA	55
13992658...	31/5/2004 ...	charlotte	bullock	F	bungaree ...	stretton pa...	wagoora	2154	charlotte.b...	BP	69
13992658...	24/9/2004 ...	esme	jardine	M	kirby place	seldon	woodville ...	3233	esme.jardi...	FEVER	39
13992658...	30/3/2004 ...	andrea	rothe	M	kinsella st...	rosetta vill...	kumbia	4305	andrea.rot...	FEVER	38
13992658...	29/10/200...	charlotte	lund	M	leichhardt ...	canberra ...	paddington	3197	charlotte.l...	DYSPEPSIA	85
13992658...	7/12/2004 ...	michael	ryan	F	abercorn c...	the willows	noble park	4076	michael.ry...	DYSPEPSIA	55
13992658...	5/5/2004 0...	clodagh	ryan	M	holden cre...	boonal	corwoa	5454	clodagh.ry...	BRONCHI...	68
13992658...	31/10/200...	imogen	warnett	M	crozier circ...	fairbank	brighton	5081	imogen.w...	GASTRITIS	86
13992658...	19/3/2004 ...	kimberley	went	M	aurora clo...	foveaux ter...	surry hills	6721	kimberley....	FEVER	86
13992658...	11/11/200...	alissa	lombardi	F	duncan str...	banksia s...	oak flats	2031	alissa.lom...	BP	96
13992658...	19/7/2004 ...	thomas	crouch	M	stapylton s...	emerald n...	jan juc	5113	thomas.cr...	GASTRITIS	6
13992658...	17/1/2004 ...	holly	cranefield	F	port jacks...	downland...	highgate	5015	holly.crane...	BP	42
13992658...	26/7/2004 ...	kane	buckmaster	M	tytherleigh ...	parks cott...	copmanhu...	3783	kane.buck...	FEVER	23
13992658...	4/2/2004 0...	jacob	morrison	M	tardent str...	ainslie ho...	campbellt...	2144	jacob.morr...	FEVER	55
13992658...	2/11/2004 ...	robbie	noble	F	jenkinson ...	gannet ho...	maryborou...	2261	robbie.nob...	BRONCHI...	40
13992658...	8/5/2004 0...	casey	beaty	F	magenta s...	the rocks	moorooka	4032	casey.beat...	FEVER	47
13992658...	29/8/2004 ...	elly	snell	F	epacris pl...	maclee	dalma	2315	elly.snell...	BP	45

(a) Original Table

Generalization replaces a value with a "less-specific but semantically consistent" value. Three types of encoding schemes have been proposed for generalization: Global recoding has the property that multiple occurrences of the same value are always replaced by the same generalized value. Regional record is also called multi-dimensional recoding (the Mondrian algorithm) which partitions the domain space into non- intersect regions and data points in the same region are represented by the region theyare in. Local recoding does not have the above constraints and allows different occurrences of the same value to be generalized differently. Generalization consists of substituting attribute values with semantically consistent but less precise values. For example, the month of birth can be replaced by the year of birth which occurs in more records, so that the identification of a specific individual is more difficult. Generalization maintains the correctness of the data at the record level but results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous dataset.

In generalization there are several recordings. The recoding that preserves the most information is "local recoding". In local recoding first tuples are grouped into buckets and then for each bucket, one replaces all values of one attribute with a generalized value, because same attribute value may be generalized differently when they appear in different buckets.

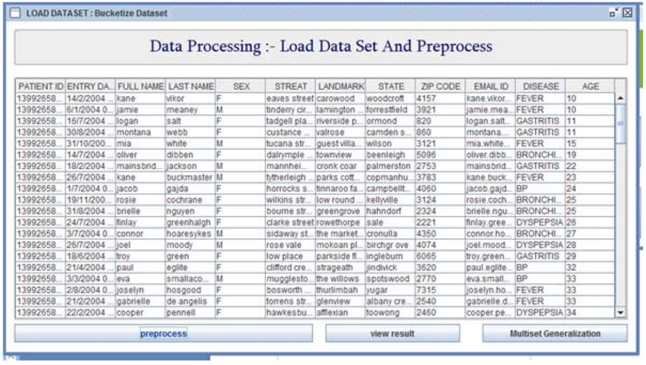

LOAD DATASET : Generalized Dataset
Data Processing :- Load Data Set And Preprocess
preprocess
view graph
BUCKETIZATION

PATIENT ID	ENTRY DA...	FULL NAME	LAST NAME	SEX	STREAT	LANDMARK	STATE	ZIP CODE	EMAIL ID	DISEASE	AGE
13992658...	14/2/2004 ...	kane	vikor	*	eaves street	carowood	woodcroft	4157**	kane.vikor...	FEVER	10 - 30
13992658...	6/1/2004 0...	jamie	meaney	*	tinderry cir...	lamington ...	forrestfield	3921**	jamie.mea...	FEVER	10 - 30
13992658...	16/7/2004 ...	logan	salt	*	tadgell pla...	riverside p...	ormond	820**	logan.salt...	GASTRITIS	10 - 30
13992658...	30/8/2004 ...	montana	webb	*	custance ...	valrose	camden s...	860**	montana....	GASTRITIS	10 - 30
13992658...	31/10/200...	mia	white	*	tucana str...	guest villa...	wilson	3121**	mia.white...	FEVER	10 - 30
13992658...	14/7/2004 ...	oliver	dibben	*	dalrymple ...	townview	beenleigh	5096**	oliver.dibb...	BRONCHI...	10 - 30
13992658...	18/2/2004 ...	mainsbrid...	jackson	*	mannhei...	cronk coar	palmerston	2753**	mainsbrid...	GASTRITIS	10 - 30
13992658...	26/7/2004 ...	kane	buckmaster	*	tytherleigh ...	parks cott...	copmanhu...	3783**	kane.buck...	FEVER	10 - 30
13992658...	1/7/2004 0...	jacob	gajda	*	horrocks s...	tinnaroo fa...	campbellt...	4060**	jacob.gajd...	BP	10 - 30
13992658...	19/11/200...	rosie	cochrane	*	wilkins str...	low round ...	kellyville	3124**	rosie.coch...	BRONCHI...	10 - 30
13992658...	31/8/2004 ...	brielle	nguyen	*	bourne str...	greengrove	hahndorf	2324**	brielle.ngu...	BRONCHI...	10 - 30
13992658...	24/7/2004 ...	finlay	greenhalgh	*	clarke street	rowethorpe	sale	2221**	finlay.gree...	DYSPEPSIA	10 - 30
13992658...	3/7/2004 0...	connor	hoaresykes	*	sidaway st...	the market...	cronulla	4350**	connor.ho...	BRONCHI...	10 - 30
13992658...	26/7/2004 ...	joel	moody	*	rose vale	mokoan pl...	birchgr ove	4074**	joel.mood...	DYSPEPSIA	10 - 30
13992658...	18/6/2004 ...	troy	green	*	low place	parkside fl...	ingleburn	6065**	troy.green...	GASTRITIS	10 - 30
13992658...	21/4/2004 ...	paul	eglite	*	clifford cre...	strageath	jindivick	3620**	paul.eglite...	BP	31 - 60
13992658...	2/8/2004 0...	joselyn	hosgood	*	bosworth ...	thurlimbah	yugar	7315**	joselyn.ho...	FEVER	31 - 60
13992658...	21/2/2004 ...	gabrielle	de angelis	*	torrens str...	glenview	albany cre...	2540**	gabrielle.d...	FEVER	31 - 60
13992658...	3/3/2004 0...	eva	smallaco...	*	mugglesto...	the willows	spotswood	2770**	eva.small...	BP	31 - 60

(b) Generalized Table

Bucketization[14,15] 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 consists of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high dimensional data. However, their approach assumes a clear separation between QIs and SAs. In addition, because the exact values of all QIs are released, membership information is disclosed. We show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. This example show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

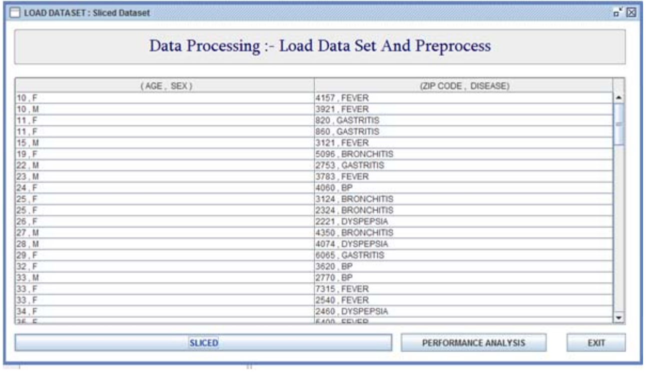
In bucketization also attributes are partitioned into columns, one column contains QI values and the other columnmcontains SA values. In bucketization, one separates the QI and SA values by randomly permuting the SA values in each bucket. In some cases we cannot determine the difference between them two. so it has one drawback for microdata publishing. It also does not prevent membership disclosure.


LOAD DATASET : Bucketize Dataset
Data Processing :- Load Data Set And Preprocess
preprocess
view result
Multiset Generalization

PATIENT ID	ENTRY DA...	FULL NAME	LAST NAME	SEX	STREAT	LANDMARK	STATE	ZIP CODE	EMAIL ID	DISEASE	AGE
13992658...	14/2/2004 ...	kane	vikor	F	eaves street	carowood	woodcroft	4157	kane.vikor...	FEVER	10
13992658...	6/1/2004 0...	jamie	meaney	M	tinderry cir...	lamington ...	forrestfield	3921	jamie.mea...	FEVER	10
13992658...	16/7/2004 ...	logan	salt	F	tadgell pla...	riverside p...	ormond	820	logan.salt...	GASTRITIS	11
13992658...	30/8/2004 ...	montana	webb	F	custance ...	valrose	camden s...	860	montana...	GASTRITIS	11
13992658...	31/10/200...	mia	white	M	tucana str...	guest villa...	wilson	3121	mia.white...	FEVER	15
13992658...	14/7/2004 ...	oliver	dibben	F	dalrymple ...	townview	beenleigh	5096	oliver.dibb...	BRONCHI...	19
13992658...	18/2/2004 ...	mainsbrid...	jackson	M	mannhei...	cronk coar	palmerston	2753	mainsbrid...	GASTRITIS	22
13992658...	26/7/2004 ...	kane	buckmaster	M	tytherleigh ...	parks cott...	copmanhu...	3783	kane.buck...	FEVER	23
13992658...	1/7/2004 0...	jacob	gajda	F	horrocks s...	tinnaroo fa...	campbellt...	4060	jacob.gajd...	BP	24
13992658...	19/11/200...	rosie	cochrane	F	wilkins str...	low round ...	kellyville	3124	rosie.coch...	BRONCHI...	25
13992658...	31/8/2004 ...	brielle	nguyen	F	bourne str...	greengrove	hahndorf	2324	brielle.ngu...	BRONCHI...	25
13992658...	24/7/2004 ...	finlay	greenhalgh	F	clarke street	rowethorpe	sale	2221	finlay.gree...	DYSPEPSIA	26
13992658...	3/7/2004 0...	connor	hoaresykes	M	sidaway st...	the market...	cronulla	4350	connor.ho...	BRONCHI...	27
13992658...	26/7/2004 ...	joel	moody	M	rose vale	mokoan pl...	birchgr ove	4074	joel.mood...	DYSPEPSIA	28
13992658...	18/6/2004 ...	troy	green	F	low place	parkside fl...	ingleburn	6065	troy.green...	GASTRITIS	29
13992658...	21/4/2004 ...	paul	eglite	F	clifford cre...	strageath	jindivick	3620	paul.eglite...	BP	32
13992658...	3/3/2004 0...	eva	smallaco...	M	mugglesto...	the willows	spotswood	2770	eva.small...	BP	33
13992658...	2/8/2004 0...	joselyn	hosgood	F	bosworth ...	thurlimbah	yugar	7315	joselyn.ho...	FEVER	33
13992658...	21/2/2004 ...	gabrielle	de angelis	F	torrens str...	glenview	albany cre...	2540	gabrielle.d...	FEVER	33
13992658...	22/2/2004 ...	cooper	pennell	F	hawkesbu...	afflexian	toowong	2460	cooper.pe...	DYSPEPSIA	34

(c) Bucketized Table

To improve the current state of the art in this paper, we introduce a novel data anonymization technique called slicing [1]. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Note that when the data set contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple, there are generally multiple matching buckets. Slicing first partitions attributes into columns. Each column contains a subset of attributes. Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. Within each bucket, values in each column are randomly permutated to break the linking between different columns. Slicing does not require the separation of those two attributes. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of data and preserves better utility. Slicing partitions the dataset both horizontally and vertically. Data slicing can also handle high-dimensional data. It provides attribute disclosure protection.


LOAD DATASET : Sliced Dataset
Data Processing :- Load Data Set And Preprocess
SLICED
PERFORMANCE ANALYSIS
EXIT

(AGE , SEX)	(ZIP CODE , DISEASE)
10 , F	4157 , FEVER
10 , M	3921 , FEVER
11 , F	820 , GASTRITIS
11 , F	860 , GASTRITIS
15 , M	3121 , FEVER
19 , F	5096 , BRONCHITIS
22 , M	2753 , GASTRITIS
23 , M	3783 , FEVER
24 , F	4060 , BP
25 , F	3124 , BRONCHITIS
25 , F	2324 , BRONCHITIS
26 , F	2221 , DYSPEPSIA
27 , M	4350 , BRONCHITIS
28 , M	4074 , DYSPEPSIA
29 , F	6065 , GASTRITIS
32 , F	3620 , BP
33 , M	2770 , BP
33 , F	7315 , FEVER
33 , F	2540 , FEVER
34 , F	2460 , DYSPEPSIA

(d) Sliced Table

VI. SLICING ALGORITHMS:

Our Algorithm of "Slicing", is presented below:

1.Load Dataset;
2.Attribute Partition And Column
3.Process Tuple Partition And Buckets
4.Slicing
5.Undergo Column Generalization
6.Do Matching Buckets
7.Duplicate An Attribute In More Than One Columns
8.End;

Our algorithm consists of three phases:

- attribute partitioning
- Column generalization.
- And tuple partitioning.

An attribute partition consists of several subsets of A, such that each attribute belongs to exactly one subset. Each subset of attributes is called a column. Specifically, let there be columnsC1, C2, . . . , Cc, Then$\mu_{\,l=1} = C_i = A$ and for any $1 \leq i1 \neq i2 \leq c$, $Ci1 \cap Ci2 = Ø$ For simplicity of discussion, consider only one sensitive attribute S. If the data contain multiple sensitive attributes, one can either consider them separately or consider their joint distribution [25]. Exactly one of the c columns contains S. Without loss of generality, let the column that contains S be the last column Cc. This column is also called the *sensitive column*. All other columns {C1, C2,…..Cc-1} contain only QI attributes.

In the second phase, tuples are generalized to satisfy some minimal frequency requirement. We want to point out that column generalization is not an indispensable phase in our algorithm. As shown by Xiao and Tao [19], bucketization provides the same level of privacy protection as generalization, with respect to attribute disclosure. Although column generalization is not a required phase, it can be useful in several aspects. First, column generalization may be required for identity/membership disclosure protection. If a column value is unique in a column (i.e., the column value appears only once in the column), a tuple with this unique column value can only have one matching bucket. This is not good for privacy protection, as in the case of generalization/bucketization where each tuple can belong to only one equivalence-class/bucket. The main problem is that this unique column value can be identifying. In this case, it would be useful to applycolumn generalization to ensure that each column value appears with at least some frequency.

In the tuple partitioning phase, tuples are partitioned into buckets, no generalization is applied to the tuples.Algo. 1 gives the description of the tuple-partition algorithm. The algorithm maintains two data structures: a queue of buckets Q & a set of sliced buckets SB. Initially, Q contains only one bucket which includes all tuples and SB is empty. For each iteration, the algorithm removes a bucket from Q and splits the bucket into two buckets [1].If the sliced table after the split satisfies l-diversity, then the algorithm puts the two buckets at the end of the queue Q. 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).The main part of the tuple-partition algorithm is to check whether a sliced table satisfies ‘*l*-diversity (line 5).

```
Algorithm tuple-partition (T, ℓ)
1. Q = {T}; SB = ∅.
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 diversity-check (T, Q ∪ {B1, B2} ∪ SB, ℓ)
6. Q = Q ∪ {B1, B2}.
7. Else SB = SB ∪ {B}.
8. Return SB.
```

(1) Tuple-partition Algorithm

Algo. 2 gives a description of the diversity-check algorithm. For each tuple t, the algorithm maintains a list of statistics L[t] about t’s matching buckets. Each element in the list L[t]contains statistics about one matching bucket B: the matching probability p(t,B) and the distribution of candidate sensitive values
D(t,B)

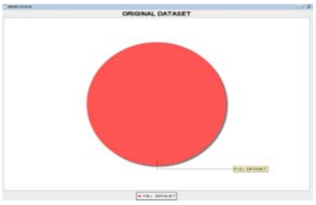
```
Algorithm diversity-check (T, T_, ℓ)
1. For each tuple t ∈ T, L[t] = ∅.
2. For each bucket B in T_
3. Record f (v) for each column value v in bucket B.
4. for each tuple t ∈ T
5. Calculate p (t, B) and find D (t, B).
6. L[t] = L[t] ∪ {p (t, B), D (t, B)}.
7. for each tuple t ∈ T
8. Calculate p (t, s) for each s based on L[t].
9. If p (t, s) ≥ 1/ℓ, return false.
10. Return true.
```

(2) The diversity-check algorithm.

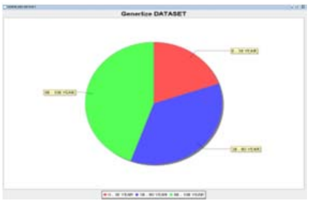
VII. Experimental Result

An important research problem is for handling highdimensional data. As per the above, Privacy Preservation for high dimensional database is important. There are two popular data anonymization technique Generalization and Bucketization. These techniques are designed for privacy preserving microdata publishing. Our Proposed work includes a slicing technique which is better than generalization and bucketization for the high dimension data sets.

Fig. a shows that the original datasets graph in which 100% means full record is shown. No any field is missing. In fig. b graph shows generalized dataset in which it divide in three groups considering any Q values. In fig, c bucketizaion process is shown in graph. In fig. d original dataset is converted in sliced dataset which is one fourth of them.



Fig(a) Original Dataset



Fig(b) Generalized Dataset

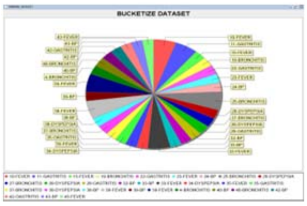
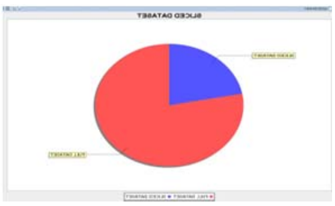


Fig (c) Bucketized Dataset



Fig(d) Slicing Dataset

VIII. CONCLUSION

In this paper, we present a new anonymization method that is data slicing for privacy preserving and data publishing. Data Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. We illustrate 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 showed some advantages of data slicing comparing with generalization and bucketization. Data slicing is a promising technique for handling high dimensional data. By partitioning attributes into columns, privacy is protected.

REFERENCES

[1] Manjusha Mirashe, Kapil Hande, “ Efficient Technique for Annonymized Microdata Preservation using Slicing", International Journal of Emerging Trends in Engineering and Development (IJTED) in Issue 5, Vol 2 in Feb2015.

[2] Li.N, Li.T, “Slicing: The new Approach for Privacy Preserving Data publishing”, IEEE Transaction on knowledge and data Engineering, vol.24, No, 3, March 2012.

[3] R. Mahesh,T . Meyyappan D “Anonymization Technique through Record Elimination to Preserve Privacy of Published Data ” IEEE Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22 2013.

[4] Ghinita.G,Tao.Y, and Kalnis.P, “OnThe Anonymization of Sparse High Dimensional Data,” Proc. IEEE 24th Int"l Conf. Data Eng. (ICDE), 2008.

[5] Amar Paul Singh, Ms. Dhanshri Parihar “A Review of Privacy Preserving Data Publishing Technique ” International Journal of Emerging Research in Management & Technology ISSN: 2278-9359 (Volume-2, Issue-6).

[6] Tamir Tassa,ArnonMazza and Aristides Gionis, “k-Concealment: An Alternative Model of k-Type Anonymity”, TRANSACTIONS ON DATA PRIVACY 5, 2012, pp189–222

[7] Qiang Wang,Zhiwei Xu and Shengzhi Qu,“An Enhanced K-Anonymity Model against Homogeneity Attack”, Journal of software,2011, Vol. 6, No.10, October 2011.

[8] Xin Jin,Mingyang Zhang,Nan Zhang and Gautam Das, “Versatile Publishing For Privacy Preservation”, 2010,KDD10,ACM.

[9] Benjamin C.M.Fung,KE Wang,Ada Wai-Chee Fu and Philip S. Yu, “Introduction to Privacy-Preserving Data Publishing Concepts and techniques, ISBN:978-1-4200-9148-9,2010

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

[11] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo,(2009), “A Survey on Privacy PreservingApproaches in Data Publishing” in the First International Workshop on Database Technology and Applications

[12] Agarwa, Srikan R., (2000) ‘‘Privacy Preserving Data Mining”, In Proc. ACM SIGMO, conferenceon management of data (SIGMOD’00), Dallas,TX,pp.439-450.

[13] Benjamin c.m, Fung, ke wang, rui chen, philips s.yu ,(2010),‘‘Privacy Preserving Data Publishing:A Survey of Recent Development ”ACM Computing surveys, Vol.42, No.4,pp.523-553.

[14] Neenu Varghese et al, Mr. Avanish Kumar Singhs “Efficient Techniques For Preserving Microdata Using Slicing” International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 1393.

[15] Sweeney L, (1996), “Replacing Personally Identifiable Information in Medical Records, the Scrub System”. Journal of the American Medical Informatics Association.

[16] Anil Prakash, Ravindar Mogili ,(2012),‘‘Privacy Preservation Measure using t-closeness withcombined l-diversity and k-anonymity”, International Journal of Advanced Research in ComputerScience and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33

[17] Charu C.Aggarwal, “A General survey of privacy preserving Data Mining Models and Algorithms”, IBM,T. J. Watson Research Centre

[18] Tiancheng Li , Jian Zhang , Ian Molloy ,(2012),“Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transaction on KDD.

[19] B.Vani, D.Jayanthi, (2013), “Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing” IJRCTT.

[20] 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.*

[21] Mohnish Patel, Prashant Richariya, Anurag Shrivastava, (2013),‘‘A review paper on Privacy-Preserving Data Mining”, Review article on Scholars Journal of Engineering and Technology(SJET) , pp.359-361