A NOVEL APPROACH TO PUBLISH DATA WITH PRIVACY PRESERVING

M. Sailakshmi[1], T Lakshmi Priya [2]

[1]Pursuing M.tech (CS) , [2]Assistant professor,

Nalanda Institute of Engineering & Technology, Siddharta Nagar,

Kantepudi (v), Sattenapalli, Guntur, Andhra Pradesh, India – 522483

ABSTRACT

Privacy preserving publishing of micro data has been studied extensively in recent years. Micro data contains records each of which contains information about an individual entity, such as a human being, a household, or an organization. Several micro data anonymization techniques have been proposed. For Ex. If a company website is considered, there the sensitive data is salary. Using Generalization security is provided by specifying salary in ranges. As salary is specifying as ranges there may be chances for losing high dimensional data. The same problem arises with suppression technique also. In Bucketization technique, the columns and rows are divided as buckets. Here sensitive data field is considered as one bucket and remaining fields as other buckets. Using Sql queries it is very easy to retrieve one field if we know all the other fields. To overcome all the problems in previous techniques, slicing technique is used. In this technique, the data is partitioned both horizontally and vertically. Here security is provided to two fields at a time. So there may be no chances for losing high dimensional data.

Keywords*: Data Loss, Micro Data, Data Publishing, Suppression Generalization, Bucketization, Slicing, Data Partition.*

I INTRODUCTION

In the recent years it's extensively studied about privacy preserving publishing of micro data. *Micro Data* contains individual records each of which having information about individual entity such as a house hold a person or an organization and so many micro data. To protect micro data so many anonymization techniques introduced, some of the important techniques are generalization k-anonymity and Bucketization in ℓ – diversity, in these two techniques attributes are divided in to 3 categories:

- ✓ some of them are identifiers, these are identified by name or some social security numbers
- ✓ Some of them are quasi attributes, this are like sex, age and address etc.
- ✓ Some of them are Sensitive Attributes these are like age and salary.

In Generalization and Bucketization the one removes the identifiers from the data after those segregate rows in to buckets. This both techniques different in second step, generalization transfer the QI – values in each buckets values so that tuples in the same bucket cannot be illustrious by their QI standards. In bucketization, one splits the SAs from the QIs by arbitrarily permuting the SA values in every bucket. The anonymized data contains of a set of buckets with permuted complex element values.

II LITERATURE SURVEY

Anonymity is the state of having one's name or identity unknown or concealed. This serves valuable social purposes and empowers individuals as against institutions by limiting inspecting threads but it is also used by wrong doers to hide their actions or avoid accountability the ability to allow anonymous access to services this avoid tracking of user's personal information and user behavior such as user location and frequency of a service usage. Suppose someone sends a file there may be information on the file that leaves a trail to the sender. Sender's data may be traced from the data logged after the file is sent.

2.1 Anonymity vs. security stability

Anonymity is a very powerful technique for protecting privacy. The inconsistence design of the Internet is particularly suitable for anonymous behavior. While anonymous actions can ensure privacy, they should not be used as the sole means for ensuring privacy as they also allow for harmful activities, such as spamming, slander, and harmful attacks without fear of reprisal. The Security threads that one should be able to detect and catch individuals conducting illegal behavior, like hacking, which tends to terrorist acts, and conducting fraud. Authentication needs for privacy should be allowed, but the ability to conduct harmful anonymous behavior without responsibility and repercussions in the name of privacy should not.

2.2 Anonymity vs. Privacy

Privacy and anonymity are not the same. There is distinction between privacy and anonymity is clearly seen in an information technology framework. The privacy corresponds to being able to send an encrypted e-mail to another recipient. Anonymity corresponds to being able to send the contents of the e-mail in plain, easily readable form but without any information that enables a reader of the message to identify the person who wrote it. Privacy is important when the contents of a message are at issue while anonymity is important when the identity of the author of a message is at issue.

2.3 Generalization

The generalization for k anonymity is loses significant on the micro data (Table 1). This is because of the following three reasons; k-anonymity suffers from the obscenity of dimensionality. In order for generalization to be better, all the records in the same bucket should be close to each other thus that generalizing the records not be lose much information. Though, in high-dimensional data, lot of data points has similar detachments with every one, forcing an excessive amount of generalization to fulfill k-anonymity flush for relative slight k's. Second, to accomplish data investigation or data mining jobs on generalized table, the data predictor has to do the constant distribution guess that each value in a generalized interval is similarly possible, as no other distribution theory can be vindicated. This meaningfully decreases the data utility of the generalized data. Third, because every element is generalized distinctly, associations between different elements are lost. While to study element associations on the generalized table, the data specialist has to adopt that each and every possible combination of element values is similarly possible. This is an integral problem of generalization that avoids effective analysis of element associations.

Age	Sex	Zip code	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 1: The Generalized Table

2.4 Bucketization

Compare with generalization bucketization is do better performance in data utilizing (Table 2). It will work in some limitations it will not prevent membership disclosure protection in the first, because bucketization circulates the QI procedures, an antagonism can discover whether a particular has a record in the circulated data or not. 87% of the persons in the United States could be inimitably identified by using only three attributes. A micro data generally encloses many other elements besides those three elements. This means that the membership information of most individuals could be conditional from the bucketized table. Second, bucketization needs a clear departure between QIs and SAs. Though, in several datasets, it is imprecise which attributes are QIs and which are SAs. Third, by extrication the sensitive element from the QI attributes, bucketization breaks the element correlations between the Quasi Interfaces and the SAs.

Age	Sex	Zip code	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

Table 2: The Bucketized Table

III PROPOSED METHODOLOGIES

We are introducing an innovative data anonymization technique calling slicing for improve the current state of the art. Slicing partitions the dataset into both vertically and horizontally. Vertical segregating is done by assemblage attributes into columns based on the relationships among the elements. Each and every column consist a subclass of attributes those are highly interrelated. Horizontal segregation is done by gathering tuples into buckets. Finally, in the each bucket, values are in each column are arbitrarily permutated to disruption the linking between different columns.

The initial idea of slicing break relationship in the cross column, but needs to sphere association with each column that will reduce the dimensionality of data and it will provide better utilization than bucketization and generalization. Slicing reserves utility because it will gather the high correlated elements together and preserves the relation between those attributes. Slicing provide security because it breaks the relationship between uncorrelated attributes, where those are not frequent and hence identifying. Note that whenever the dataset encloses one SA and QIs, bucketization will break their relationship; on the other hand slicing, can gather some Quasi Interfaces attributes with the SA, preservative element correlations with the delicate attribute.

Finally, we are conducting wide workload experimentations. Our results endorse that slicing conserves much well data utility than generalization. In our workloads including the composite attribute, slicing is also more effective than bucketization. In some cataloguing experiments, slicing shows the best performance than using the original data. Our trials also show the boundaries of bucketization protection in membership disclosure and slicing remedies these limitations.

3.1 Slicing

We propose a novel approach called slicing with some example, how it will provide efficient security on micro data than bucketization and generalization. Slicing is its capability to handle high-dimensional data. By dividing attributes into columns then slicing reduce the measurements of the data. Then each of which column of the table can be viewed as a sub-table with a lesser dimensionality. The slicing is also not similar from the approach of publishing multiple independent sub-tables in that these sub-tables are associated by the buckets in slicing.

(Age,Sex)	(Zipcode,Disease)
(22,M)	(47905,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.)

Table 3: The Sliced Table

The table displays the Generalized Table which consisting of micro data tables and their anonymization versions using various anonymization techniques. The three Quasi Interface attributes are such as Sex, Age, Zip code, etc. and Disease the sensitive attribute. A generalized table it will satisfies 4-anonymity is displayed in Table 1, a bucketized table that fulfills 2-diversity is displays in Table 2, the each attribute value is replaced in generalized table with the multi set of values in bucket is shown in the Table 2, and the two sliced tables are displayed in the Table 3. The Slicing that will first partitions element into columns. Each and every column consist a subset of attributes. This is vertically segregates the table. Slicing also segregates the each record into buckets, each and every bucket contains the information about record, and this will horizontally segregate the table data.

3.2 Formalization of Slicing

Assume T is the micro data table. T having d attributes:

$d = \{A_1, A_2, A_3\ldots A_d\}$ and their element domains are $\{D [A_1], D [A_2], D [A_3]\ldots D [A_d]\}$. A row $t \in T$ can be represented as t = (t [A1], t [A2], t [Ad]) where t [Ai] ($1 \leq i \leq d$) is the Ai value of t.

***Definition 1*:** (Attribute columns and partitions). An element partition having of several subsets of A, such that each attribute fits to exactly one subset. Each subset of attributes is called as columns. Specifically, let there be c columns, C1, C2. Cc, then

∪c i=1Ci = A and for any 1 ≤ i1 6= i2 ≤ c, Ci1 ∩ Ci2 = ∅.

For easiness of conversation, we think only one sensitive attribute S. If in case the data comprises multiple sensitive attributes, one could whichever consider them separately or consider their joint distribution. Accurately one of the c columns contains S. Without loss of overview, let the column that having S be the last column Cc. This column is moreover called the sensitive column. All the other columns {C1, C2. Cc−1} having only QI attributes.

Definition 2: (Tuples partition and buckets). A tuple partition contains of few subsets of T, such that every tuple pertaining to exactly one subset. Every subset of tuples is called as a bucket. Specifically, let there is a b buckets, B1, B2. Bb, then

∪b i=1Bi = T and for any 1 ≤ i1 6= i2 ≤ b, Bi1 ∩ Bi2 =∅.

***Definition 3*:** (Slicing) it specifies a micro data table T, a slicing of T is specified by an attribute partition and a tuple partition. Often times, slicing also comprises column generalization.

Definition 4: (Column Generalization). Displayed a micro data table T and a column Ci = {Ai1,Ai2, . . . ,Aij}, the column generalization for Ci is defined as a set of no overlying j dimensional areas that completely cover D[Ai1] × D[Ai2] × . . . × D[Aij]. A column generalization maps every value of Ci to the area in which the value is enclosed.

3.3 Slicing Vs Generalization

In present days few recoding methods are available for generalization in local systems these recoding techniques will preserve more information in the local systems. In the local recoding systems, they first cluster the tuples in buckets, after that each bucket one value attribute is replaces with generalized values. This recoding is local, because this generalization may be done differently in another tuples, even though if the same values are appears in the different bucket.

We now describing that slicing preserving more information compare with the local recoding technique. Assume uses same tuples partition is used. We will reach this by showing Slicing is better than the following enrichment is better than local coding approach. Instead of using a generalized value to replace more precise attribute values, one use the multi set of precise values in each and every bucket. The multi set of particular values delivers lot of information about the spreading of values in every attribute than the generalized interval. Therefore, using multi set of correct values reserves more information than generalization.

Another essential benefit of slicing is its capability to handlebar high-dimensional data. By segregating the attributes into columns, slicing decreases the dimensionality of the data. Every column in the table can be watched as a sub-table with a minor dimensionality. Slicing is different from the method of publishing multiple independent sub-tables in that these sub-tables are connected by the buckets in slicing.

3.4 Slicing Vs Bucketization

When we compare slicing with bucketization, we initially note that bucketization can be watched as a distinctive case of slicing, where there are accurately two columns: one column encloses only the SA, and the other comprises all the QIs. The benefit of doing slicing on bucketization can be understood follows. First, by segregating attributes into more than two columns, slicing could be used for prevent membership disclosure. Second, dissimilar bucketization, which entails a clear separation of QI features and the sensitive attribute, slicing is used without such a parting. For dataset such as the survey data, one often cannot clearly separate QIs from SAs because there is no single external public database that one can use to determine which attributes the adversary already knows. In this case slicing can be useful for such data. Finally, by allowing a column to contain both some QI attributes and the most important attribute, attribute correlations between the sensitive attribute and the QI attributes are conserved.

3.5 Privacy Issues

While publishing micro data there are 3 types of privacy disclosure issues.

Membership Disclosure: Whenever the dataset needs to be published, it needs to select from large population and the selection criteria sensitive data like a particular disease values. One desires to prevent opponents from access whether one's record is included in the published dataset or not.

Identity Disclosure: Which arises when a discrete is linked to a specific record in the released table? In few situations, one needs to keep from identity disclosure when the opponent is indefinite of membership. In this situation, defense against membership disclosure supports protect against identity disclosure. In other conditions, some opponent may previously know that an entity's record is in the distributed dataset, in which situation, membership disclosure protection whichever does not apply or is inadequate.

Attribute Disclosure: This arises whenever new information about some individuals is exposed, i.e., the unconfined data will make it possible to suppose the attributes of an individual exactly than it could be probable before to release. Similar to the case of identity disclosure, we need to deliberate opponents who already know the membership information. Identity disclosure signs to attribute disclosure. A discrete is re-identified, once there is distinctiveness disclosure and the corresponding sensitive value is revealed. Attribute disclosure can arise with or without identity disclosure, e.g., whenever the sensitive values of all matching tuples are the same.

3.6 Algorithm Implementation

We are presenting an effective slicing algorithm to accomplish ℓ-diverse slicing. IN given micro data table T and two parameters c and ℓ, the algorithm calculates the sliced table that contains c columns and fulfills the privacy requirement of ℓ-diversity.
Our algorithm contains three phases: column generalization, attribute partitioning and tuple partitioning. Now we define the three phases.

Attribute Partitioning

Our algorithm segregates the attributes so that highly interrelated attributes are arranged in the same column. This is somewhat good for both activities like utility and privacy. In the case of data utility, gathering highly correlated attributes preserves the correlations among those elements. In the case of privacy, in the relationship of uncorrelated attributes shows higher identification threats than the relationship of highly correlated attributes cause the overtone of uncorrelated attribute values is more less numerous and therefore more recognizable. Therefore, it is somewhat better to disruption the relationships between uncorrelated elements, while defensive privacy. In this section, we first analyze the connections between sets of attributes and then group attributes based on their correlations.

Column Generalization

In the second section, tuples are generalized to fulfill some common occurrence requirement. We wish to concentrate on column generalization is not a crucial phase in our algorithm. As shown by Tao and Xiao, bucketization tuple partition (T, ℓ) algorithm

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.

It is providing the same level of privacy protection as generalization, with the attribute disclosure.

Tuple Partitioning

In the tuple separating section, tuples are divided into buckets. For tuple partition we have to change the Mondrian algorithm dissimilar Mondrian k-anonymity, no generalization is performed to the tuples; we have to use Mondrian for the determination of segregating tuples into buckets. The main aim of the tuple-partition algorithm is to validate whether a sliced table satisfies ℓ-diversity (line 5).

Diversity check (T, T_, ℓ) algorithm

1. For each tuple t ∈ T, L[t] = ∅.
2. For each buckets 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] ∪ {hp (t, B), D (t, B) i}.
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.

Tuple t, the algorithm conserves a list of statistics L[t] about its matching buckets. Each element in the list L[t] comprises statistics about one matching bucket B: the matching possibility p (t, B) and the distribution of candidate sensitive values D (t, B).

IV CONCLUSIONS

A novel data anonymization technique called *slicing* to providing privacy micro data publishing. Slicing partitions the data set both vertically and horizontally. Multi-dimensional space adds a new dimension for partitioning. Slicing overwhelms the limitations of bucketization and generalization and preserves better efficacy while securing against privacy threats. We demonstrate how to usage slicing to stop attributes disclosure and membership disclosure. Our experiment displays that slicing preserves best data utility than generalization and is more efficient than bucketization in workloads concerning the sensitive attribute. The general technique proposed by this work is that before doing data anonymization, somebody can analyze the characteristics of data and they will use these characteristics in data anonymization. The basis is that one can design best data anonymization techniques when we know the data better.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy "Slicing: A New Approach to Privacy Preserving Data Publishing" on IEEE Transactions on Knowledge and Data Engineering.

[2] Neha v. Mogre, sulbha patil "slicing: an approach for privacy preservation in High-dimensional data using anonymization Technique" on Proceedings of Fifth IRAJ International Conference.

[3] Sathish.R, Silambarashi.G, Saranya.P, Santhosh Kumar.B "A New Approach For Collaborative Data Publishing Using Slicing And M-Privacy" on International Journal of Innovative Research in Computer and Communication Engineering

[4] Data Mining Concepts, Tasks And Techniques Author-S N Sivanandam , S Sumathi.

[5] B.Santhosh Kumar ,"Privacy Preserving Data Publishing For Multiple Sensitive Attributes".

AUTHOR PROFILE

M. Sailakshmi is currently pursuing M.Tech in the Department of Computer Science & Engineering, from Nalanda Institute of Engineering & Technology (NIET), siddharth Nagar, Kantepudi(V), Sattenapalli (M), Guntur (D), Andhra Pradesh , Affiliated to JNTU-KAKINADA.

T Lakshmi Priya working as Assistant Professor at Nalanda Institute of Engineering & Technology (NIET), siddharth Nagar, Kantepudi(V), Sattenapalli (M), Guntur (D), Andhra Pradesh , Affiliated to JNTU-KAKINADA.