Implementing A Hybrid Slicing Approach for Privacy Preserving Data Publishing

Vikas Baisane, N. D. Kamble

Abstract*: Several techniques of anonymity, such as generalization and disruption, have been designed to protect privacy from the publication of micro-data. Recent work has shown that generalization loses much information, especially for high dimensional data. Bucketization, on the other hand, does not prevent the disclosure of membership and does not apply to data that do not have a clear separation between quasi-identifiable attributes and sensitive attributes. In this paper, we present a new technique called overlapping slice, which divides the data horizontally and vertically. We show that the overlap section preserves a better data utility than generalization and can be used for the protection of belonging to belonging. Another important advantage of the overlap slice is that it can handle large data storage.

***Keywords*: Privacy preservation, data anonymization, data publishing, data security**

I. INTRODUCTION

Data anonymization[5] is a technology that converts plain text into a form that cannot be read by human beings. The technique of anonymizing data for privacy, data publishing has received a lot of attention in recent years. Detailed data (also called microdata) contains information about a person, a household or an organization. The most popular anonymization techniques are generalization and scarcity. [1] There are a number of attributes in each record that can be categorized as 1) Identifiers such as Name or Social Security Number are the attributes that can be uniquely identified to individuals. (2) some attributes may be sensitive attributes (SA) such as illness and salary, and (3) some may be quasi-identifiers (IQs) such as postal code, age and gender whose values, they are taken together, can potentially identify an individual. Data is considered anonymous, even when it is associated with pointer or pedigree values that direct the user to the system, the record and the original value.

We consider a new type of "initial attack" in collusion of data providers who can use their own data records (a subset of the global data) in addition to the external knowledge base to infer data records contributed by d other data providers. The paper addresses this new threat and makes several contributions. First, we introduce the notion of m-privacy, which ensures that anonymity data satisfies a given confidentiality constraint against any collusive data provider group up to m. Second, we present heuristic algorithms exploiting the monotony of the confidentiality constraint equivalence group and adaptive control techniques to effectively verify the deprivation of power given a set of records. Finally, we present an anonymity algorithm using a data provider with appropriate strategies to verify the deprivation of power in order to ensure an efficient and useful use of anonymity data. Experiments on real-life datasets suggest that our approach allows better and more usefulness and efficiency than existing and basic algorithms while providing a guarantee of confidentiality.

Revised Version Manuscript Received on September 22, 2017.

Vikas Baisane, D.N. Patel College of Engineering, Shahada, Dist: Nandurbar (Maharashtra), India. E-mail: vikasbaisane2009@gamil.com

Mr. N. D. Kamble, Asst. Prof., Shreeyash College of Engineering and Technology, Education: ME (CSE), Aurangabad (Maharashtra), India.

II. RELATED WORK

Generalization for the anonymity [1] of k-anonymity results in a considerable amount of information, especially for high-dimensional data. Bucketisation [2]does not prevent the disclosure of membership. Because degradation publishes IQ [3][4] values in their original forms, an opponent can know if an individual has a record in published data or not. Shrinking goals requires a clear separation between QI and SAs. However, in many datasets it is not clear which attributes are IQs and which are SAs. We assume that data providers are semi-honest, commonly used in computing distributed computing. They may attempt to infer additional information about data from other providers by analyzing the data received on anonymous basis. Several micro data anonymity techniques have been proposed. The most popular ones are generalization for k-anonymity[3] and bucketization[3][4][5] for ‘-diversity. In both approaches, attributes are partitioned into three categories:

A. Some attributes are identifiers that can uniquely identify an individual, such as a Social Security Number or Name;

B. Some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other databases available to the public) and which together can potentially identify an individual, eg date of birth, sex, and Postal Code;

C. Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Illness and Wage.

In both generalization and bucketization, one first removes identifiers[5][6] from the data and then partitions tuples into buckets. The two techniques differ in the next step. The generalization transforms the IQ values in each cube into "less specific but semantically coherent" [7].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

values so that the tuples in the same cube can not be distinguished by their IQ values[8]. In bucketization, one separates the SAs from the QIs by randomly exchanging the SA values in each cube.

III. PROPOSED SYSTEM

We introduce a new data anonymity technique called overlapping slice to improve the current state of the art. Overlapping sliced partitions[7][8] allows data to be defined both vertically and horizontally. Vertical partitioning is done by grouping the attributes in columns according to the correlations between the attributes. Each column contains a subset of highly correlated attributes. Horizontal partitioning[8][9] is done by grouping tuples into buckets. Finally, within each bucket, the values of each column are randomly (or sorted) to break the link between different columns. The basic idea of overlapping slicing is to break the transverse association columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves a better utility than the generalization and the disturbance. Slicing overlay preserves utility because it aggregates highly correlated attributes and maintains the correlations between these attributes. The overlap summit protects privacy because it breaks the associations between uncorrelated attributes, which is infrequent and therefore identifiable. Note that when the dataset contains IQs and SAs, the lapse must interrupt their correlation; the overlapping blocking, on the other hand, can group certain QI attributes with the SA, preserving the attribute correlations with the sensitive attribute. The key intuition that overlapping slicing provides privacy protection is that the overlapping slicing process ensures that for any tuple there are usually multiple buckets.

We consider the collaborative parameter for publishing data with data partitioned horizontally across several data providers, each providing a subset of Ti records. As a special case, a data provider could be the owner of the data itself which contributes its own records. This is a very common scenario in social networks and recommendation systems. Our goal is to publish an anonymity view of the embedded data so that a data recipient, including data providers, will not compromise the confidentiality of individual records provided by other parties.

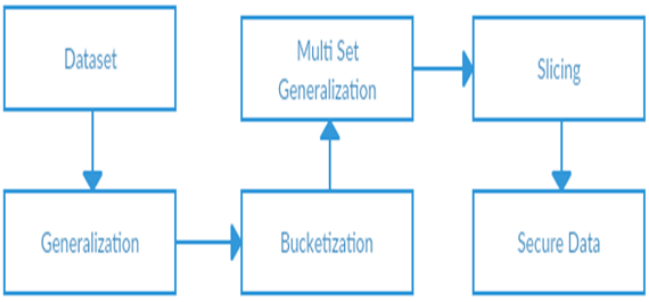


Figure 1 Architectural Flow Diagram

A. Attribute Partitioning

This algorithm partitions the attributes so that the highly correlated attributes are in the same column. This is good for utility and privacy. In terms of data utility, the grouping of highly correlated attributes preserves the correlations between these attributes. In terms of confidentiality, the association of uncorrelated attributes presents higher identification risks than the association of strongly correlated attributes because associations of uncorrelated attribute values are much less frequent and therefore more identifiable.

B. Tuple Partitioning

The algorithm maintains two data structures: 1) a bucket queue Q and 2) a set of sliced buckets SB. Initially, Q contains a single bucket that includes all tuples and SB is empty. For each iteration, the algorithm removes a bucket from Q and divides the bucket into two buckets. If the sliced table after the division satisfies the diversity l, then the algorithm puts the two buckets at the end of the queue Q Otherwise, we can no longer divide the bucket and the algorithm puts the bucket in SB. When Q becomes empty, we compute the table in slices. The whole of the sliced buckets is SB.

C. Slicing

The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple there are usually multiple buckets. Given a tuple t hv1; v2; . . . ; vci, where c is the number of columns and vi is the value of the i th column, a bucket is a corresponding bucket for t if and only if for each i (1 _ i _ c) vi appears at least once in the i th column of the bucket. Every bucket containing the original tuple is a matching bucket. At the same time, a matching bucket may have to contain other tuples, each containing some but not all vi.

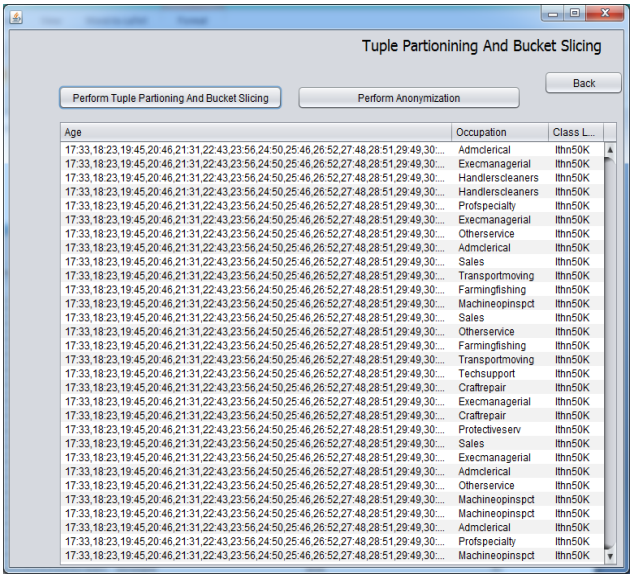
Given a microdata table T, a slicing of T is given by an attribute partition and a tuple partition. For example, Tables 1e and 1f are two sliced tables. In Table 1e, the attribute partition is {{Age}, {Sex}, {Zipcode}, {Disease}} and the tuple partition is {{t1; t2; t3; t4}, {t5; t6; t7; t8}}. In Table 1f, the attribute partition is {{Age, Sex}, {Zipcode, Disease}} and the tuple partition is {{t1; t2; t3; t4}, {t5; t6; t7; t8}}.

The three IQ attributes are {age, sex, postal code}, and the sensitive attribute SA is a disease. A generalized table that satisfies anonymity 4 is shown in Table 1 (b), the data in the zed table of the satisfying 2-diversity hole are presented in Table 1 (c), a generalized table where each attribute value is The values in the bucket are shown in Table 1 (d), and two sliced tables are shown in Table 1 (e) and 1 (f). The first partition attributes in columns. Each column contains a subset of attributes. For example, the sliced table in Table 1 (f) contains 2 columns: the first column contains {Age, Sex} and the second column contains {zip code, disease}. The sliced table shown in Table 1 (e) contains 4 columns, each column containing exactly one attribute. Slicing is also partitions tuples in buckets. Each bucket contains a subset of tuples. This horizontally partition the table. For example, the two sliced tables in Table 1 (e) and Table 1 (f) contain 2 buckets, each containing 4 tuples.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

In each bucket, the values in each column are randomly swapped to break the connection between different columns.

For example, in the first bucket of the sliced table shown in Table 1 (f), the values {(22, M), (22, F), (33, F), (52, F)} are permuted (47906, Influenza), (47905, Influenza), (47905, bronchitis)} are randomly swapped so that the link between the two columns in a bucket is hidden.



D. Generalization

The generalization module performs an anonymity process 2. In the generalization approach, we use the data of identifiers and Quasi Identifiers. Here, the age of the attribute is Identifiers, and the gender is Quasi Identifiers. The generalization data can be retrieved from an original data. The data set data is stored in two buckets.

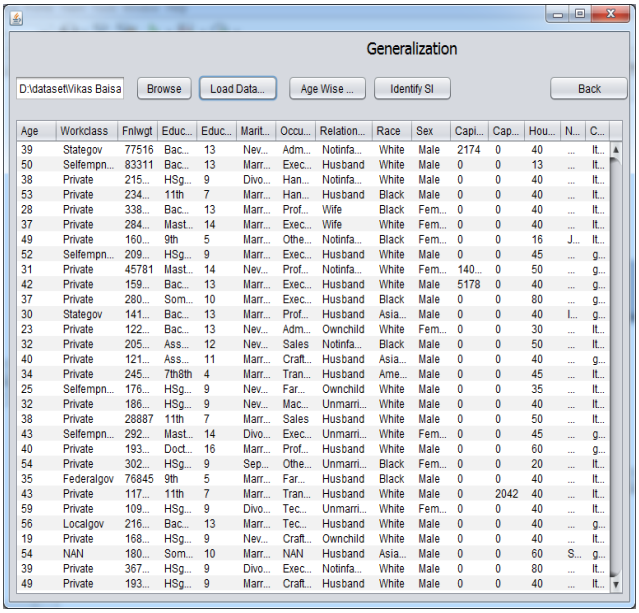
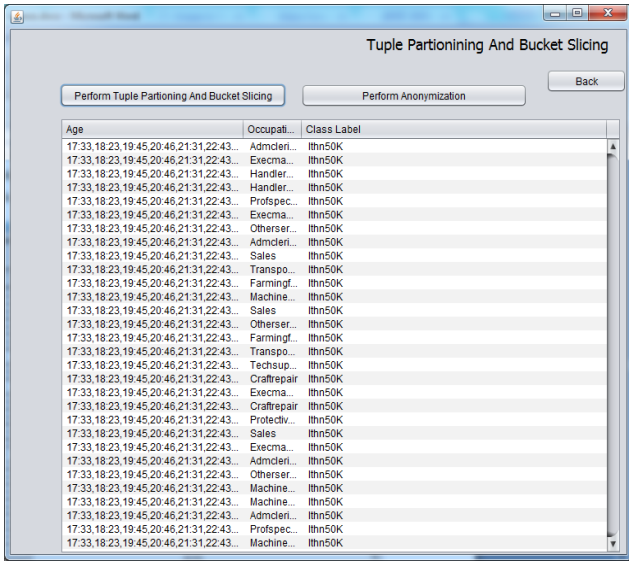


Figure 2 Generalization

E. Bucketization

The Bucketization module can perform a 2-diversity process. In the generalization approach, we use the Quasi Identifiers. Here, the attribute work class assigns. The decontamination data can be retrieved from an original data. The data set data is stored in two buckets.



F. Multi-Set Generalization

The multi-set generalization module performs an anonymity process 2. In the multi-set generalization approach we use the identification data and Quasi Identifiers. Here, the age of the attribute is Identifiers, and the gender, work class are Quasi Identifiers. The multi-set generalization data can be retrieved from an original datum. The data set data is stored in two buckets.

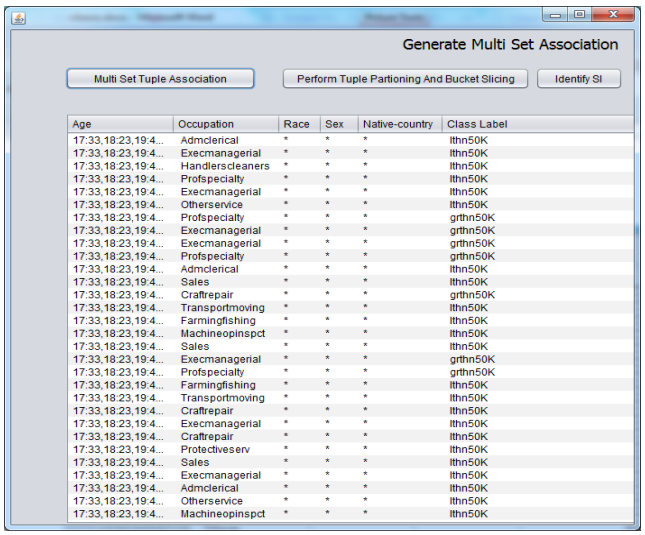


Figure. 3 Multi Set Association

G. Overlapping Slicing

Overlapping sliced partitions allows data to be defined both vertically and horizontally. The overlap section preserves a better utility of data than generalization and can be used for the protection of belonging to belonging. Here we use the following sub-modules.

- Attribute partition and Columns
- Tuple Partition and Buckets
- Overlapping slicing
- Column Generalization
- Matching Buckets

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

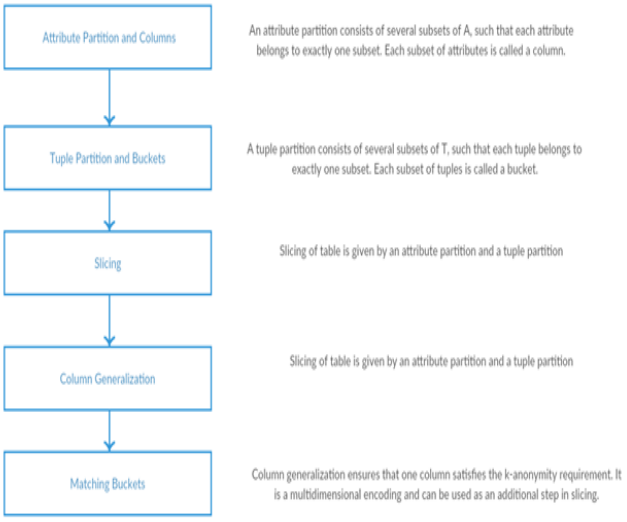


Figure 3 Data Flow Diagram

IV. ALGORITHMS

A. Algorithm for Tuple Partition

Algorithm tuple-partition(T, 0
Q = {T}; SB = O.
while Q is not empty
remove the first bucket B from Q; Q = Q — {B}.
create list of non empty buckets
obtain non empty list and extract successive buckets
split B into two buckets B and B2
if diversity-check(T,Q U {B1, B2} U SB, E)
Q = Q u
{-Bi,B2}
else
SB = SB U {B}
return SB

B. Algorithm for Diversity Check

Algorithm diversity-check(T, T*, f)
for each tuple t E T, L[t] = 0.
for each bucket B in T*
record 1(v) for each column value v in bucket B.
calculate mid, median, min and max values
median= list.size()/2;
min = list.get(first)
Mid = median
max =list.size()-1
for each tuple t E T
calculate p(t, B) and find D(t, B).
L[t] = L[t]U {(p(t, B), D(t, B))}.
for each tuple t E T
calculate p(t, s) for each s based on L[t].
if p(t, s) > 1/f, return false.
return true.

V. EXPERIMENTAL RESULTS

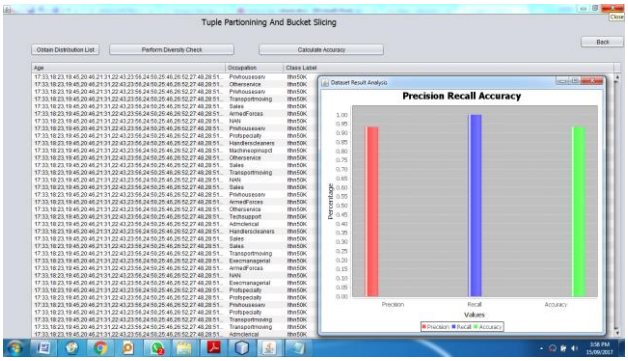
In our experiments, we obtain two sets of data from the adult dataset. The first dataset is the "OCC-7" dataset, which consists of seven attributes: IQ = {Age, class of work, education, marital status, race, sex} and S = Occupation. The second set of data is the "OCC-15" dataset, which includes the 15 attributes and the sensitive attribute is S = Occupation. Note that we do not use Salary as the sensible attribute because the salary has only two values {> = 50K; <50Kg}, which means that even 2-diversity is not achievable when the sensible attribute is Salary. Also note that in the protection of disclosure of membership, we do not differentiate between IQs and SAs.

In result analysis with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance.

In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are:

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp+tn/tp+tn+fp+fn$



VI. COMPUTATIONAL EFFICIENCY

We compare the slicing to the generalization and the reduction of the consumption in terms of calculation efficiency. We repair l = 5 and modify the cardinality of the data (i.e. the number of records) and the dimensionality of the data (i.e. the number of attributes). Figure 6a shows the calculation time as a function of the data cardinality where the data size is set to 1 = 5 (i.e. we use (subsets) of the OCC-15 dataset).

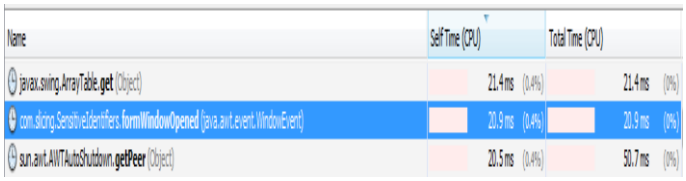
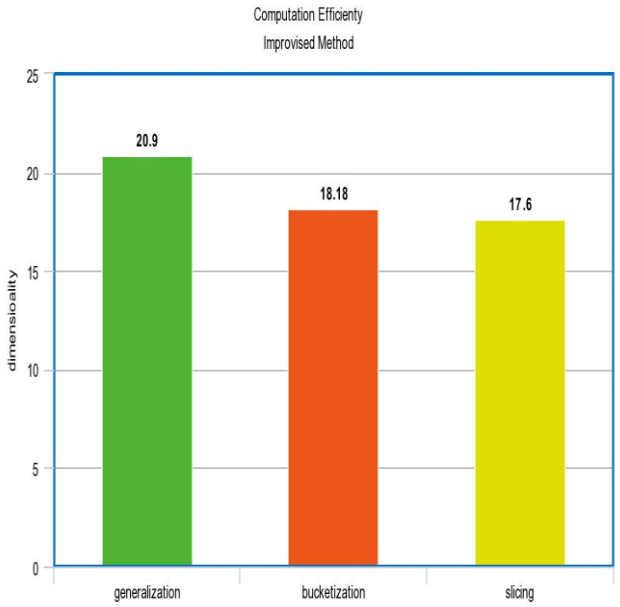


Figure 5. Java Profiling for Generalization, Bucketization and Slicing Methods

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.



VII. ATTACKS BY EXTERNAL DATA RECIPIENT USING ANONYMITY DATA

A data recipient, for example, P0, could be an attacker and attempts to infer additional information about the records using the published data (T *) and some background knowledge (BK) such as the externally available public data.

VIII. ATTACKS BY DATA PROVIDERS USING ANONYMITY DATA AND THEIR OWN DATA

Each data provider, such as P1 in FIG. 1, can also use the anonymity data T * and its own data (T1) to infer additional information on other records. Compared to the attack of the external recipient in the first attack scenario, each provider has additional data knowledge of its own records, which can help the attack.

IX. FUTURE SCOPE AND CONCLUSION

Reduce the limits of generalization and discoloration and preserve better utility while protecting against threats of confidentiality. Slicing prevents disclosure of attribute and disclosure of belonging. Slicing preserves a better utility of data than generalization and is more effective than reducing workloads involving the sensitive attribute.

We plan to cut out where each attribute is exactly in a column. An extension is the notion of superimposed break, which doubles an attribute in more than one column. Our experiments show that random grouping is not very efficient. We plan to design more efficient tuple grouping algorithms. Another direction is to design data mining tasks using anonymous data [6] computed by various anonymization techniques. Slicing protects privacy by breaking the association of uncorrelated attributes and preserving the utility of data by preserving the association between highly correlated attributes. Another important advantage of slicing is that it can handle high-dimensional data.

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" Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.
2. V. Ciriani, S. De Capitani di Vimercati,S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).
3. Latanya Sweeney. K-anonymity: "a model for protecting privacy". International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.
4. 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
5. Benjamin C. M. Fung, Ke Wang, AdaWai-Chee Fu, and Philip S. Yu, "Privacy Preserving Data Publishing Concepts and Techniques" ,Data mining and knowledge discovery series (2010).
6. Neha V. Mogre, Girish Agarwal, PragatiPatil: "A Review on Data Anonymization Technique For DataPublishing" Proc. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December-2012 ISSN: 2278-0181
7. N. Li, T. Li, and S. Venkatasubramanian,"t-Closeness: Privacy Beyond k-Anonymity and 'l-Diversity,"Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.
8. A. Machanavajjhala, D. Kifer, J. Gehrke,and M. Venkitasubramaniam. "l-diversity: Privacy beyond kanonymity". InICDE, 2006.
9. D. Martin, D. Kifer, A. Machanavajjhala,J. Gehrke, and J. Halpern. "Worst-case background knowledge for privacy-preserving data publishing". In ICDE, 2007.
10. G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'lConf. Data Eng. (ICDE), pp. 715-724, 2008.
11. R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization,"in Proc. of ICDE, 2005, pp. 217–228.
12. K. LeFevre, D. J. DeWitt, and R.Ramakrishnan, "Incognito: Efficient Full-domain k-Anonymity," in Proc. of ACM SIGMOD, 2005, pp. 49–60.
13. K. LeFevre, D. J. DeWitt, and R.Ramakrishnan, "Mondrian Multidimensional k-Anonymity," in Proc. of ICDE, 2006.
14. Gabriel Ghinita, Member IEEE, PanosKalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol.23no. 2) pp. 161-174.
15. 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

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.