ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(21), DOI: 10.17485/ijst/2016/v9i21/95142, June 2016* ISSN (Online) : 0974-5645

Slicing+: An Efficient Privacy Preserving Data Publishing

M. Nithya[1,2*] and T. Sheela[3]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai - 600119, Tamil Nadu, India; nithya.cse@sairam.edu.in

[2]Department of Computer Science and Engineering Sri Sairam Engineering College, Chennai - 600044, Tamil Nadu, India

[3]Department of Information Technology, Sri Sairam Engineering College, Chennai - 600044, Tamil Nadu, India; hod.it@sairam.edu.in

Abstract

Objectives: Privacy and accuracy are always trade off factors in the field of data publishing. Ideally both the factors are considered critical for data handling. Privacy loss and accuracy loss need to be maintained low as possible for an efficient data handling system. Authors have come up with various data publishing techniques aiming to achieve balance between these 2 factors. Generalization, Bucketization and Slicing are well known techniques among the list. Unfortunately they have their own limitation in handling privacy and accuracy. Generalization suffers in handling high dimensional data thus experiencing higher utility loss. Bucketization lacks data privacy where parting sensitive and quasi identifier attributes is a challenge. Slicing on the other hand though offers better privacy and accuracy, there is always scope to improve data correlation aiming in reducing utility loss. This paper explains a new technique called Slicing+ which handles privacy and accuracy factors effectively. This new Slicing+ technique looks promising as it offers flexibility for data publisher to decide on how the data need to be published. Data publisher can tune the Slicing+ technique to get data published with better privacy than accuracy or the other way. Algorithms for the two cases are derived and realized using Orange tool. This paper explains analysis done for the first bucket tuples. As an improvement aspect, similar analysis can be done for other buckets and all the bucket tuples merged and reconstructed for complete analysis. This analysis is applied in the medical records. This hybrid slicing technique is rated against Privacy loss and Utility gain factors. Experimental results are analyzed to justify the performance of Slicing+ technique.

Keywords: Accuracy, Data Mining, Privacy, Publishing, Slicing

1. Introduction

Data publishing is considered as a critical stage in data analysis system. Publishing sensitive data might lead to individual privacy breach. Predictive rules and techniques[1] can help in predicting privacy information easily. Thus data anonymization becomes a requirement to avoid sensitive data leakage. Anonymization techniques like Generalization[2,3], Bucketization[4–6] and Slicing are well known which handle data anonymization in their own way. In general, these techniques manage in manipulating the original data to avoid sensitive data made available for data analysts. In this course of data manipulation, there are always possibilities of data utilization going down. Utilization loss becoming predominant shall directly affect the accuracy of data analysis. In few occasions the analysis results go completely wrong finally unable to solve the very purpose of data mining and publishing. In general there is a strong assumption that privacy and accuracy are trade off features[7], practically impossible to achieve both. This paper disagrees with the assumption and explains both privacy and accuracy can be achieved by transforming the algorithm based on the need[8]. Open source Orange data mining tool is used to design a new algorithm called Slicing+ which is the successor of Slicing technique explained in[9]. Further two cases are discussed where in the first case privacy of data is focused and in the second case accuracy of data is focused. In both the cases the other trade off factor still offers promising results thus giving this Slicing+ technique a new dimension. Initial part of this paper will detail on merits and demerits of Generalization, Bucketization and Slicing techniques.

** Author for correspondence*

Further the paper would explain how Slicing+ is realized using Orange Tool which is one of the open source data mining tool developed in Python by Faculty of Computer science in University of Ljubljana.

2. Data Analysis

Any source data shall have identifiers which can uniquely identify individual (Name, SSO), Quasi Identifiers (Age, Sex) which are available for the analyst and finally sensitive data (Disease, Salary) whose privacy need to be secured. Medical records from Hospital, salary records from Company are considered as sensitive data which are prone to security issues and attacks[10]. These data when leaked out could be a threat to individual privacy. These data could be of any data type, volume and size. The data source can be manipulated with certain level of privacy maintained and released to certain group of people. In parallel another group of people might receive manipulated data with different degree of privacy. If both the data variants are somehow accessible by an intruder then there is always a possibility to compare both the data variants and exploit the privacy factor. Further data analysis results should always respect analysis requirement. In few occasions maintaining data privacy is expected than accuracy of data. In other cases accuracy of data is mandate. Thus data publishing technique should be flexible for generating reports as per need.

3. Inspiration for Slicing+

3.1 Generalization

Generalization for k-anonymity has higher data utility loss for high dimensional data[11–14]. This is due to the fact that data Generalization in a bucket requires data closeness. If the tuples distances are far apart then Generalization could be a challenge. In general data closeness cannot be expected for high dimensional data. Further uniform distribution assumption is required for tuples falling in a bucket which further impacts data utility. Adding to the above utility issues, Generalization is done separately for each attribute thus impacting data correlation between attribute columns.

3.2 Bucketization

Bucketization technique has better data utility when compared to Generalization but has serious privacy concerns. Membership disclosure cannot be prevented[15] in this technique as all the quasi identifiers are published in their natural state. This can create opportunity for intruder to decode the identifier information based on the quasi identifier relation. In general certain quasi identifier combination when undistorted can be used to find the identifier information easily. Another drawback is this technique requires perfect demarcation between quasi identifiers and sensitive attributes. In most cases there is always confusion in identifying quasi identifiers vs. sensitive attributes. Thus this concern adds to Bucketization drawback. Finally attribute correlation gets affected as this technique needs parting of quasi identifier and sensitive attribute.

3.3 Slicing

Slicing has an upper hand with respect to Generalization and Bucketization. In this technique the source data table is handled both vertically and horizontally. In this technique the source data table is divided column wise. This division brings certain quasi identifiers together on one side (vertical X) and the other with a combination of quasi identifier and sensitive attribute (vertical Y). Further the data table is bucketed and random shuffling is executed on the second part (vertical Y) of bucketed tuples. This random shuffling has certain probability of disclosing sensitive attributes when certain tuples after shuffling retains the same old tuple position. This creates an opportunity for realizing an efficient shuffling technique. Further care should be taken when there are repeating sensitive attributes with ineffective shuffling resulting in greater probability of membership disclosure.

4. Slicing+ Technique

4.1 Methodology

Considering the drawbacks of Generalization, Bucketization and Slicing there is a need to create an improvement in publishing technique. Slicing+ technique is designed considering the above drawbacks. In this technique, number of buckets are identified based on the count of sensitive data variants. Source data are then column wise segregated into 2 parts with sensitive data and quasi identifiers separated. Further, sensitive data parts are shuffled to guarantee privacy (Case 1) or sensitive data copied to quasi identifiers to guarantee accuracy (Case 2).

4.2 Algorithm

Case 1: Preserving Privacy

N: Dataset, n (N) = No. of tuples in N; N has 5 attributes: $N = \{N_V, N_W, N_X, N_Y, N_Z\}$; $N_1 = \{N_W, N_X, N_Y, N_Z\}$

Partitioning the data set N_1 in to number of buckets based on attribute values in N_Z.

$N_1 = B_1 \cup B_2 \cup B_3 \ldots\ldots B_M$ where M = Number of buckets with 2- diversity in N_Z for N_1.

Vertical $X = B_W B_Y$, Vertical $Y = B_X B_Z$. i.e $N_1 = B_W B_Y \cup B_X B_Z$, $B_W B_Y \cap B_X B_Z = \emptyset$.

$B_W B_Y = \cup B_{Wi} B_{Yi}$, $B_X B_Z = \cup B_{Xi} B_{Zi}$, where i = 1 to M

$B_W B_Y = B_{W1} B_{Y1} \cup B_{W2} B_{Y2} \cup B_{W3} B_{Y3} \ldots\ldots \cup B_{WM} B_{YM}$

$B_X B_Z = B_{X1} B_{Z1} \cup B_{X2} B_{Z2} \cup B_{X3} B_{Z3} \ldots\ldots\ldots\ldots \cup B_{XM} B_{ZM}$

Quasi identifiers: V = Name; W = Location; X = Age; Y = Gender and Sensitive attribute: Z = Disease.

For each $B_X B_Z$, perform random shuffling based on sensitive attribute.

For tuples with identical sensitive data, perform shuffling again and merge shuffled vertical Y with X.

Case 2: Retaining Accuracy

N: Dataset, n (N) = No. of tuples in N; N has 5 attributes: $N = \{N_V, N_W, N_X, N_Y, N_Z\}$; $N_1 = \{N_W, N_X, N_Y, N_Z\}$

Partitioning the data set N_1 in to number of buckets based on attribute values in N_Z.

$N_1 = B_1 \cup B_2 \cup B_3 \ldots\ldots B_M$ where M = Number of buckets with 2-diversity in N_Z for N_1.

Vertical $X = B_W B_Y$, Vertical $Y = B_X B_Z$. i.e $N_1 = B_W B_Y \cup B_X B_Z$, $B_W B_Y \cap B_X B_Z = \emptyset$.

$B_W B_Y = \cup B_{Wi} B_{Yi}$, $B_X B_Z = \cup B_{Xi} B_{Zi}$, where i = 1 to M.

$B_W B_Y = B_{W1} B_{Y1} \cup B_{W2} B_{Y2} \cup B_{W3} B_{Y3} \ldots\ldots \cup B_{WM} B_{YM}$

$B_X B_Z = B_{X1} B_{Z1} \cup B_{X2} B_{Z2} \cup B_{X3} B_{Z3} \ldots\ldots\ldots\ldots \cup B_{XM} B_{ZM}$

Quasi identifiers: V = Name; W = Location; X = Age; Y = Gender and Sensitive attribute: Z = Disease.

For each $B_X B_Z$, perform random shuffling based on sensitive attribute.

$N_2 = B_{Wi} B_{Yi} \cup B_{Zi}$ and $N_3 = N_2 \cup$ shuffled $B_{Xi} B_{Zi}$

4.3 Working Procedure

Step 1: Source data table to be published is first classified into identifiers, quasi identifiers and sensitive attributes. Since attribute "disease" is a confidential data for an individual which can reveal personal information when linked with other attributes it is considered as sensitive data.

Step 2: Sensitive attribute data is taken into consideration to determine number of buckets. Numbers of sensitive data variants are counted. This count is divided by required l-diversity number. the result determines number of buckets into which the source table shall fall within. Each bucket shall have number of tuples based on expected l-diversity[16]. Horizontal partition is realized in this step .

Step 3: The source data table is next divided into columns. This division brings certain quasi identifiers together on one side (vertical X) and the other with a combination of quasi identifier and sensitive attribute (vertical Y).

Step 4: Vertical component Y which is a combination of quasi identifier and sensitive attribute is shuffled such that post shuffling none of the Vertical component Y tuple falls into its original position.

Step 5: For tuples having identical sensitive data, next level of shuffling is done with respect to the quasi identifier linked to the sensitive attribute in the vertical component Y. This step guarantees better privacy.

Step 6: The original sensitive data are copied to the X component which improves data utility but on the other side could impact data privacy.

Step 7: The final manipulated data are published by the data analyst. Slicing+ offers better data utility while retaining data privacy. Step 6 is ignored if data utility can be traded off with respect to privacy.

4.4 Experimental Analysis

Both cases of Slicing+ technique are experimented. In Case 1, privacy is considered critical and thus Step 6 of algorithm is ignored. Case 2 experiments the same source data assuming data accuracy is critical compared to data privacy. Thus Step 6 of Slicing+ algorithm is included to bring better data correlation and data utilization. For both the cases initial steps are common and thus explained in general. Figure 1 shows 40 medical records which are sensitive in nature are considered for publishing. Unique identifier data "Name" is removed from source table for further processing. Location", "Gender" and "Age" are considered as quasi identifiers and "Disease" as sensitive attribute. Combination of the quasi identifiers in the table could easily reveal the sensitive information.

Name	Location	Gender	Age	Disease
Raj	Chennai	M	64	Malaria
kani	Mumbai	F	25	Diabetics
ramu	Chennai	M	13	Flu
Rani	Indore	F	24	Malaria
bubu	Delhi	M	56	Cancer
bana	Indore	F	27	Flu
Mona	Mumbai	F	12	Diabetics
monu	Mumbai	M	28	Malaria
temu	Chennai	M	79	Typhoid
Peter	Mysore	M	21	Cancer
Sumi	Chennai	F	45	Flu
Sona	Mysore	F	36	Malaria
Preti	Mumbai	F	32	Malaria
Giri	Delhi	M	34	Cancer
Gaja	Mysore	M	89	Flu
Kajal	Delhi	F	79	Diabetics
Priya	Chennai	F	39	Flu
Prem	Indore	M	40	Malaria
Tamil	Chennai	M	41	Diabetics
Puna	Delhi	F	52	Cancer
Baji	Mumbai	M	43	Flu
Vani	Chennai	F	44	Cancer
Loha	Mumbai	F	09	Malaria
Diva	Mysore	F	45	Typhoid
Meru	Mumbai	F	46	Cancer
Papu	Mumbai	M	36	Flu
Meto	Delhi	F	48	Malaria
Tupi	Chennai	F	49	Diabetics
tepu	Mumbai	M	50	Cancer
Qumi	Mysore	F	51	Flu
Kitu	Chennai	M	22	Malaria
Ani	Indore	F	53	Diabetics
Mina	Mumbai	F	76	Cancer
Joti	Chennai	F	46	Flu
Kumar	Mysore	M	56	Malaria
Hima	Mumbai	F	57	Diabetics
Teja	Delhi	M	09	Cancer
Kiran	Mysore	M	59	Flu
kapu	Mumbai	F	60	Malaria
Suga	Mumbai	F	12	Typhoid

Figure 1. Source data.

To avoid the combination of quasi identifier, they have to be delinked. 6 variants of information are identified in the sensitive attribute column. 2 diversity per bucket is considered for this experiment. Thus we arrive at original data bucketed into 3 buckets with 2 diversity. First data bucket is considered for analysis in Orange tool.

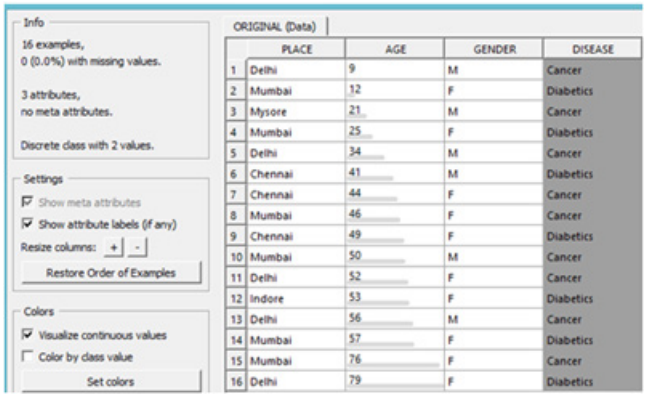


Figure 2. Bucket 1 data.

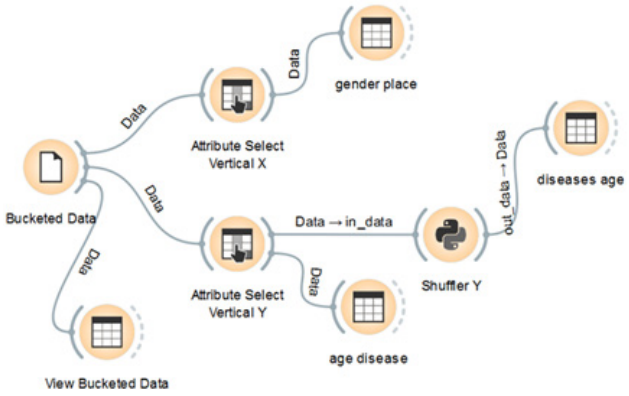


Figure 3. Slicing+ in Orange Tool.

Figure 2 shows the bucket 1 data with only 2 variants of sensitive attribute. 16 tuples are grouped in this bucket. Figure 3 explains the programming of Slicing+ algorithm. Bucketized data are further split into columns with "Place" and "Gender" data selected under Vertical X component. Rest of "age" and "disease" columns are selected under Vertical Y component.

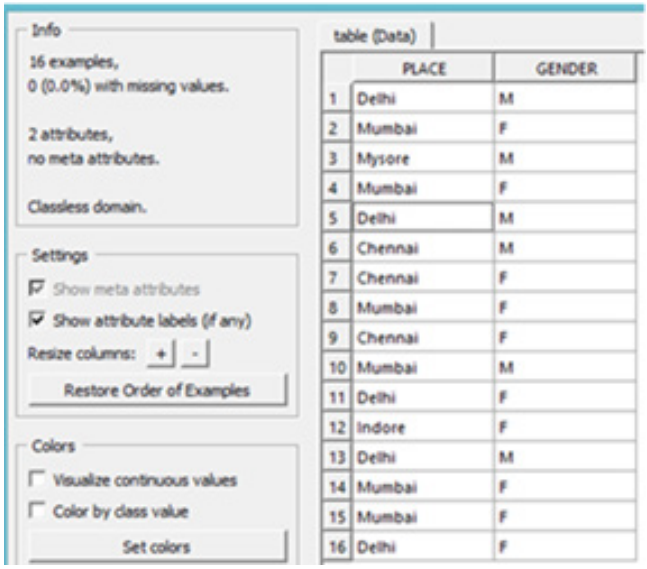


Figure 4. Vertical X component partitioned data.

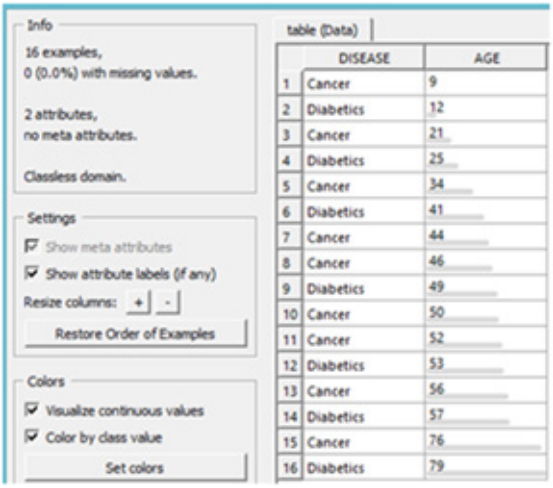


Figure 5. Vertical Y component partitioned data.

Figure 4 shows the vertical X component partitioned data. "Attribute Select Vertical X" function selects "place" and "gender" quasi identifier columns into it. Figure 5 shows the vertical Y component partitioned data. "Attribute Select Vertical Y" function selects "age" and "disease" columns into it. Vertical Y component partitioned data is further processed using Shuffler Y function. This function is programmed to shuffle both the age and disease columns together in such a way, post shuffling none of the Vertical component Y tuple falls into its original position. For tuples having identical sensitive data, next level of shuffling is done with respect to the "age" quasi identifier linked to the sensitive attribute "disease" in the vertical component Y. This step guarantees better privacy.

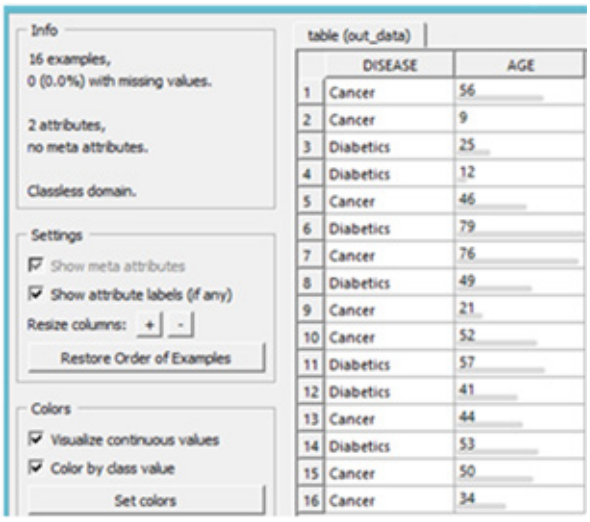


Figure 6. Post shuffled status of Vertical Y component.

Figure 7. Reconstruction of data bucket.

Figure 6 shows the post shuffled status of Vertical Y component. Comparing tables in Figure 5 and 6, there is no existence of "age" and "disease" row combination as per original state. This condition guarantees maximum data privacy. Algorithm step 6 is ignored in this case. Figure 7 shows how original Bucketized data is reconstructed. Both vertical X and Y components are merged using merge data function.

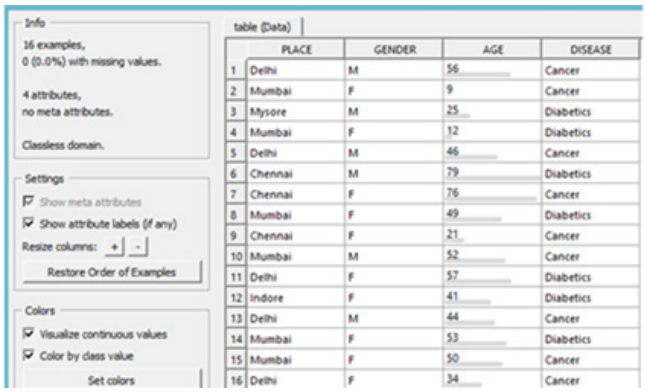


Figure 8. Reconstructed data bucket for Case 1.

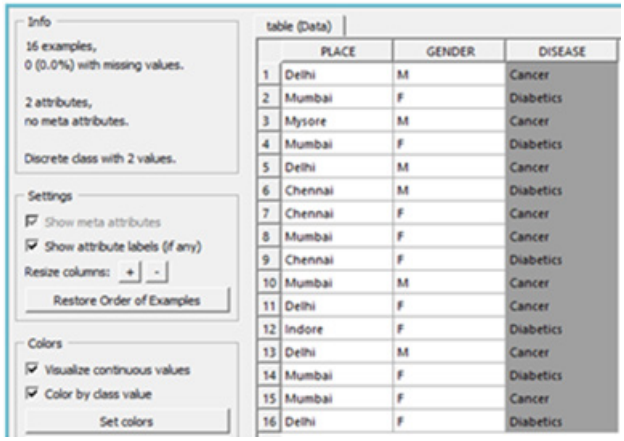
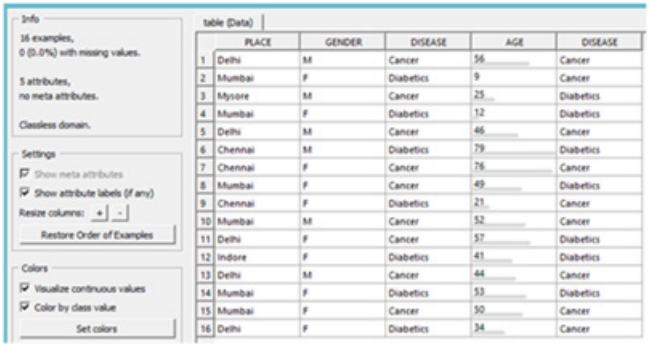


Figure 9. Original disease column in X component.

Data publishing doesn't require merging both vertical X and Y component. This reconstruction is done for

analysis purpose only. Figure 8 shows the reconstructed data bucket for Case 1. Privacy of data is guaranteed whereas data utility might become a challenge as correlations of data columns is lost while slicing data vertically. To overcome this drawback, the experiment is repeated again. In this case Step 6 is executed. The original sensitive data "disease" is copied to the X component. Figure 9 shows the addition of original disease column in the X component. This addition could improve data utility but on the other side could impact data privacy. Figure 10 shows the reconstructed data bucket for Case 2. Utility of data is guaranteed whereas data privacy might become a challenge as sensitive attribute is reflected in both the vertical components.



	PLACE	GENDER	DISEASE	AGE	DISEASE
1	Delhi	M	Cancer	56	Cancer
2	Mumbai	F	Diabetics	9	Cancer
3	Mysore	M	Cancer	25	Diabetics
4	Mumbai	F	Diabetics	12	Diabetics
5	Delhi	M	Cancer	46	Cancer
6	Chennai	M	Diabetics	79	Diabetics
7	Chennai	F	Cancer	76	Cancer
8	Mumbai	F	Cancer	49	Diabetics
9	Chennai	F	Diabetics	21	Cancer
10	Mumbai	M	Cancer	52	Cancer
11	Delhi	F	Cancer	57	Diabetics
12	Indore	F	Diabetics	41	Diabetics
13	Delhi	M	Cancer	44	Cancer
14	Mumbai	F	Diabetics	53	Diabetics
15	Mumbai	F	Cancer	50	Cancer
16	Delhi	F	Diabetics	34	Cancer

Figure 10. Figure 10. Reconstructed data bucket for Case 2

5. Experimental Results

Experimental results are analyzed using scatter plots. Sensitive data "Disease" is plotted with respect to place, age and sex.

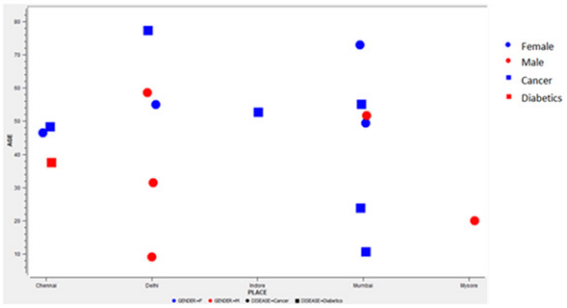


Figure 11. Scatter plot for source data.

Figure 11 shows scatter plot of source data. Privacy loss is measured with a simple formula: Each tuple attribute is given a weightage and weightage sum of the tuple is calculated in the source table and Post Slicing+ table. Weightage sum is compared for both the tables. If same weightage values appear in both the tables, then their tuple attribute combination is verified. In this case privacy is leaked. This test is conducted with source table and Case 1 Sliced+ table. Similar test is conducted with source table and Case 2 Sliced+ table. It is evident that privacy loss is more for Case 2 Sliced+ table when compared to Case 1 Sliced+ table. Accuracy is measured by comparing the scatter plots of source table and Sliced+ table. Disease patterns are analyzed to see if the same results can be arrived with the data from source table and Sliced+ table. Similar the distance pattern, better accuracy can be arrived.

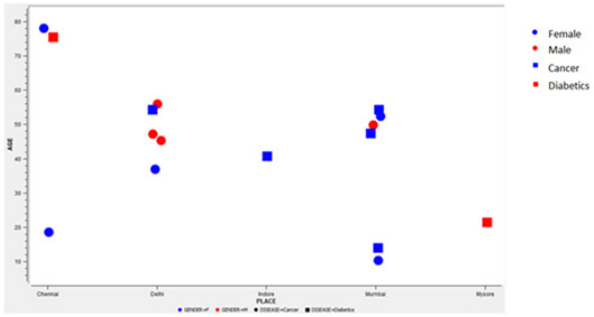


Figure 12. Scatter plot for Case 1 reconstructed data.

Figure 12 shows scatter plot of reconstructed data. Analysing the disease pattern reveals, there is huge variation in pattern positioning. Thus accuracy shall be a challenge for Case 1 results. Case 2 when reconstructed has 2 columns of sensitive data appearing in both vertical component X and Y. The intruder might analyse the table by having one of the sensitive column at a time to predict sensitive attribute. The intruder might repeat the analysis by having the other sensitive column next. Thus we have 2 possibilities in Case 2. First analysis is done by retaining the original sensitive column and removing the shuffled sensitive column. This becomes Case 2a.

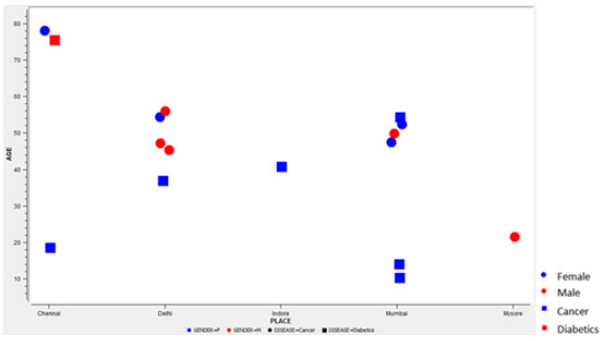


Figure 13. Scatter plot for Case 2a reconstructed data.

Figure 13 shows scatter plot of Case 2a reconstructed data. Analyzing the disease pattern reveals, there is lesser variation in pattern positioning. Thus accuracy is improved for Case 2a when compared to Case 1 results. Further analysis is done by retaining the shuffled sensitive column and removing the original sensitive column. This becomes Case 2b.

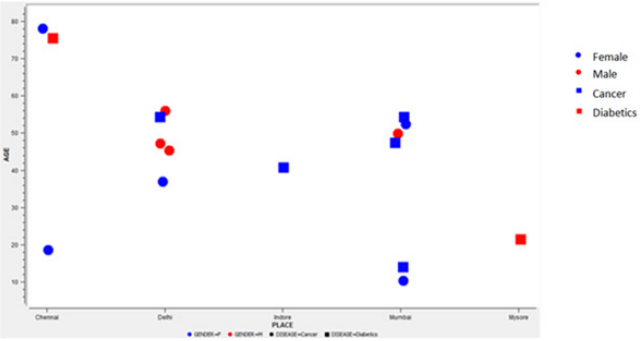


Figure 14. Scatter plot for Case 2b reconstructed data.

Figure 14 shows scatter plot of Case 2b reconstructed data. Analysing the disease pattern reveals, there is lesser variation in pattern positioning similar to Case 2a results. Accuracy is improved for Case 2b when compared to Case 1 results. To conclude, Case 2 has better accuracy and data utility when compared to Case 1.

6. Graph Interpretation

Figure 11 shows the scattered plot for source data. This is considered as the base plot for comparing Case 1 and Case 2 reconstructed data. Accuracy is attained when similar graphical pattern is achieved between 2 graph sets. Comparing Figure 11 and 12, it is evident that there is variation in the graphic pattern. Also the distance between the plot points are not near. Thus accuracy is not achieved in Case 1. Almost similar graphic pattern is achieved between Figure 11 and Figures 13 and 14. Thus accuracy is achieved in Case 2. Privacy is achieved when there is no data leakage. Data is leaked if same weightage sum is achieved for the tuples in source table with respect to Case 1 or Case 2 reconstructed table. Case 2 has same weightage sum for most of the tuples thus loosing privacy when compared to Case 1. Thus Case 1 has better privacy and Case 2 with better accuracy.

7. Conclusions

This new Slicing+ technique looks promising as it offers flexibility for data publisher to decide on how the data is required to be published. Data publisher can tune the Slicing+ technique to get data published with better privacy than accuracy or the other way. This paper explains analysis done for the first bucket tuples. Similar analysis can be done for other buckets and all the bucket tuples need to merged and reconstructed for complete analysis. This provides room for future work. Another area to concentrate shall be on the shuffling technique which plays a vital role in determining data privacy.

8. References

1. Balakrishnan V, Shakouri MR, Hoodeh H. Integrating association rules to predict retinopathy. Maejo International Journal of Science and Technology. 2012 Sep: 6(03):334–43.
2. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowledge and Data Eng. 2001 Nov; 13(6):1010–27.
3. Sweeney L. k-Anonymity: A model for protecting privacy. Int'l J Knowledge-Based Systems. 2002 Oct; 10(5):557–70.
4. Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. Proc Int'l Conf Very Large Data Bases; 2006 Sep. p. 139–50.
5. Martin DJ, Halpern JY. Worst-case background knowledge for data publishing. Int'l Conf Data Eng; 2007 May. p. 126–35.
6. Koudas N, Yu T, Zhang Q. Aggregate query answering on anonymized tables. Int'l Conf Data Eng; Istanbul. 2007 Apr 15-20. p. 116–25.
7. Brickell J, Shmatikov V. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. KDD; 2008 Aug. p. 70–8.
8. Li T, Li N. Tradeoff between privacy and utility in data publishing. Int'l Conf Knowledge Discovery and Mining; 2009 Jul. p. 517–26.
9. Li T, Molloy I. Slicing: A new approach for privacy preserving data publishing. Trans on Knowledge and Data Eng. 2012 Mar; 24(3):561–74.
10. Thummavet P, Vasupongayya S. Privacy-preserving emergency access control for health records. Maejo International Journal of Science and Technology. 2015 Apr; 9(01):108–20.
11. Aggarwa Cl. On k-Anonymity and the curse of dimensionality, Proc Int'l Conf Very Large Data Bases (VLDB); 2005 Aug. p. 901–9.
12. Kifer D, Gehrke J. Injecting utility into anonymized data

sets. Proc ACM Int'l Conf Management of Data; 2006 Jun. p. 217–28.
13. Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. Proc Int'l Conf Very Large Data Bases; 2006. p. 139–50.
14. Manikandan G, Sairam N, Sharmili S, Venkatakrishnan S. Achieving privacy in data mining using normalization. Indian Journal of Science and Technology. 2013 Apr; 6(4):1–5.
15. Nergiz ME, Clifton C. Hiding the presence of individuals from shared databases. Int'l Conf Management of Data; 2007 Jun. p. 665–76.
16. Machanavajjhala A, Venkitasubramaniam M. L'-Diversity: Privacy beyond k-anonymity. Proc Int'l Conf Data Eng; Atlanta, GA. USA 2006 Apr 3-7. p. 24.