A HYBRID APPROACH FOR PRIVACY PRESERVING DATA MINING

Tatshini G.[1], Shaili Jha[1] and R. Devika[2]
[1]B.Tech Computer Science Engineering, SASTRA University, Thanjavur, India
[2]Computer Science Engineering, School of Computing, SASTRA University, Thanjavur, India
E-Mail: shaili_jha@yahoo.co.in

ABSTRACT

Privacy preservation has become a major issue during data publishing. The simplest solution is not to disclose the information. This would be of no use since it will hinder the process of data analysis. Instead, the data can first be modified so that it can guarantee privacy and, at the same time, it retains sufficient utility and can be released to other parties safely. The existing system uses a hybrid approach of sampling and generalization for data modification. An alternative system using cross over and mutation of genetic algorithm for data modification, without the use of a trusted third party is proposed. In addition, shearing technique and double pass matrix based encryption technique are implemented to improve the confidentiality of the data.

Keywords: privacy preservation, crossover, matrix based encryption, shearing, privacy preserving data publishing (PPDP).

1. INTRODUCTION

Data publishing is the process of making data, collected by various institutions such as hospitals, financial institutions, government etc. public, so that it can be used to find useful patterns for the purpose of data mining. But the growing concern with this is the violation of the individual's privacy. When seen in isolation, the individual's identity may not be found out. But this can be linked with other publicly available documents like voter list and the sensitive information of an individual may be compromised. One approach to overcome this is to apply transformation techniques to the sensitive attributes in a manner that the data published will be useful for analysis while still protecting the privacy. A term closely related to this is privacy preserving data mining (PPDM) that refers to development of data mining techniques that have privacy concerns incorporated. These involve data hiding, rule hiding etc. Data perturbation is a specific class of data hiding techniques.

In a typical scenario, there are three actors involved. They are: Record Owner, Data Publisher and Data Recipient. The interactions between them are depicted below:

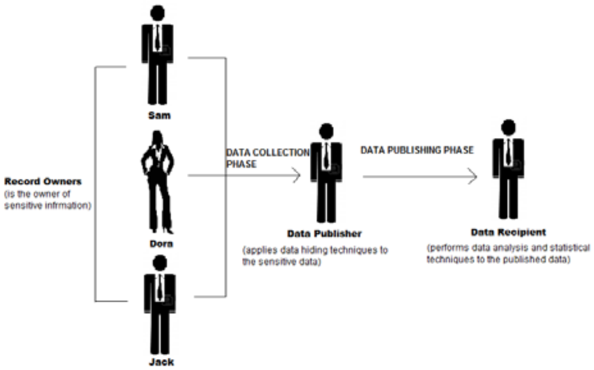

Sam
Dora
Jack
Record Owners
(is the owner of sensitive infrmation)
DATA COLLECTION PHASE
Data Publisher
(applies data hiding techniques to the sensitive data)
DATA PUBLISHING PHASE
Data Recipient
(performs data analysis and statistical techniques to the published data)

Figure-1. Representation of PPDM.

The existing system uses a hybrid approach of sampling and generalization. The original data set is first sampled. Then, a formula involving the expected and observed confidence is formed. Thus the sampled and generalized data set is published for the purpose of analysis. The proposed system for secure/impregnable release of data, which is also useful for statistical purposes, is to perturb the sensitive data by applying the techniques of Crossover (and Mutation if required), Double pass Matrix based Encryption and Shearing to the fields (numerical data only) specified by the user.

2. METHODOLOGY

The data to be perturbed are given to the Data publisher as depicted. The Record owner is given the freedom to apply the three proposed techniques namely, Crossover, Double pass Matrix based Encryption and Shearing, to the numerical attributes. Crossover is applied to a quasi identifier that has low tolerance to deviation, Double pass Matrix based Encryption to the unique identifier and Shearing to any other numerical quasi identifier. Double pass Matrix based Encryption and Shearing techniques are applied not to whole value of the attribute as such. Instead, a part of the value of the attribute that might reveal sensitive information is taken into consideration. This ensures that data analysis can still be carried out in an efficient manner.

2.1 Crossover

Crossover is one of the operators applied in genetic algorithms. It is a twostep process. After reproduction, members of the resulting population are mated at random. These are then converted into binary strings. After this, an integer position is chosen at random and the bits at that position are swapped. This gives rise to two new strings.

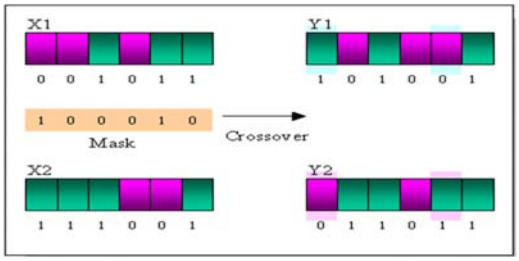


Figure-2. A typical example for Crossover.

Mutation is another operator of genetic algorithm. In this, a position is chosen at random in the string and its bit value is changed. They are depicted below:

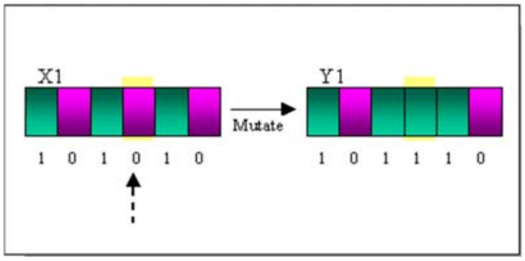


Figure-3. A typical example of Mutation.

A combination of these two techniques is adopted in the proposed system. Two attribute values (field chosen by the user to which crossover is to be applied) are selected in a sequential manner and then converted into binary values. Then the bits are paired up. A number is generated and the bits in that particular position are swapped. Mutation is applied in the following three cases:

a) When the number of records are odd in number, mutation alone is applied to the first record

b) When the bits chosen for crossover have the same value and crossing over will not make any change to the original data

c) When the value of the attribute exceeds the threshold. In such a case, the higher order bits are mutated (unlike the other two cases wherein lower order bits are taken into consideration) until the attribute's value is brought back into the specified range

2.2 Double pass matrix based encryption

Encryption is a cryptographic technique in which the message (referred to as the plaintext) is encoded using an algorithm, converting it into an indecipherable cipher text. The encryption algorithm typically involves the usage of an encryption key that defines how the message is to be encoded.

The steps followed in Double pass matrix based encryption are as follows:

a) The sensitive information is first extracted from the specific record and stored in an array

b) A number is generated at random (range of random number chosen in a way that it does not exceed the size of the database) and similarly the sensitive information of the record indicated by the random number is extracted and stored in another array

c) The matrix to be used for the process of encryption is generated next. It is a 10 x n matrix ('n' depends on the number of digits of the attribute decided on them being the sensitive information) generated in the following manner:

- The sensitive information is split up into individual digits
- A row is generated based on the individual digit, in a serial manner, by incrementing it by one, until all the digits from 0-9 are covered.
- For example: If the digit extracted is 6, then the row generated would be 6 7 8 9 0 1 2 3 4 5
- The process is repeated n times

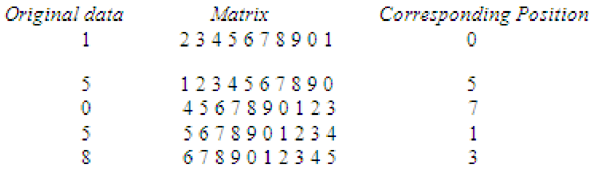
d) The position of the digit to be encrypted is found in the particular row. This value is used to encrypt the original data. This is shown in the following example:

Consider the following banking database:

S. No.	Name	Age	Gender	Account number	Pin code
1.	X	23	M	1798000015058	613401
2.	Y	25	F	1798000016330	500016
3.	Z	34	M	1798000067451	625006
4.	A	40	F	1798000040001	600028
5.	B	27	M	1798000032897	530032
6.	C	55	M	1798000021456	613004

Assume the account number is to be encrypted and the last five significant digits are taken as the significant digits. Then, another account number is taken at random and its significant digits are used as a key for encrypting the account number in question. This is done in the following format:

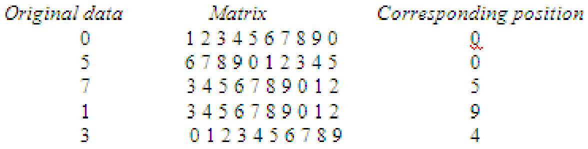
Account number to be encrypted: 1798000015058 Significant digits: 15058
Random number generated: 6
Account number chosen as key: 1798000021456 Significant digits: 21456



Original data	*Matrix*	*Corresponding Position*
1	2 3 4 5 6 7 8 9 0 1	0
5	1 2 3 4 5 6 7 8 9 0	5
0	4 5 6 7 8 9 0 1 2 3	7
5	5 6 7 8 9 0 1 2 3 4	1
8	6 7 8 9 0 1 2 3 4 5	3

e) The entire process is repeated twice to ensure that the original sensitive data cannot be traced back. Hence the name double pass matrix based encryption. Using the same example as reference, the second pass is shown below:

Account number to be encrypted: 1798000005713 Significant digits: 05713
Random number generated: 2
Account number chosen as key: 1798000016330 Significant digits: 16330



Original data	*Matrix*	*Corresponding position*
0	1 2 3 4 5 6 7 8 9 0	0
5	6 7 8 9 0 1 2 3 4 5	0
7	3 4 5 6 7 8 9 0 1 2	5
1	3 4 5 6 7 8 9 0 1 2	9
3	0 1 2 3 4 5 6 7 8 9	4

Now the account number which will replace 1798000015058 will be 1798000000594.

f) Finally, the original data is replaced by the doubly encrypted data

2.3 Shearing

Shearing is one of the graphics-based transformation techniques. In this, distortion of the object happens in a way that it appears as if the object were comprised of internal layers that had been caused to slide over each other. It may deform the shape of the object along either one of the axes: x-axis, y-axis or both.

Matrix form of shearing along X axis in 3D:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ Sh_{zx} & Sh_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Matrix form of shearing along Y axis in 3D:

$$\begin{bmatrix} 1 & Sh_{xy} & Sh_{xz} & 0 \\ Sh_{yx} & 1 & Sh_{yz} & 0 \\ Sh_{zx} & Sh_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

A similar approach is adopted in the proposed system. The formula used for shearing is:

X`=X+ (Shearing_factor x X)

Where X` is the modified data and X is the original data. Shearing_factor is the random number generated in the range [2-9]
The steps are as follows:

a) The specified digits (which contain sensitive information) are extracted from the field and stored in a variable

b) A random number is generated between [2-9]

c) This is multiplied with the extracted number and added to the original data and its average is taken

d) Modulo operation with 10^n (where n denotes the number of digits extracted from the original data) is performed for normalization in cases where in the number of digits exceeds that of the original data

e) The original data is replaced with this modified data

3. EXPERIMENTAL RESULTS

A User Interface has been created to facilitate the application of the proposed technique on the desired database.

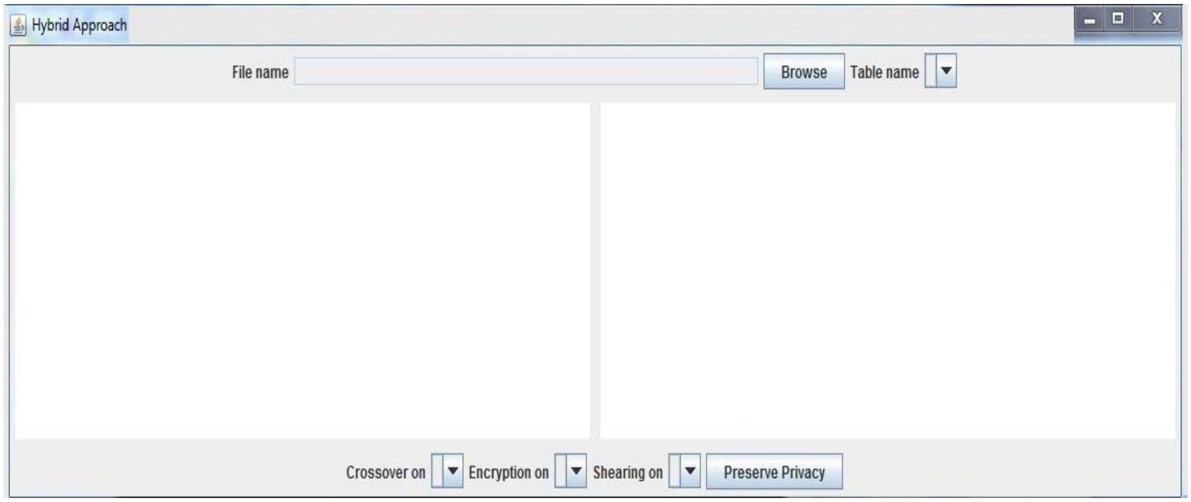
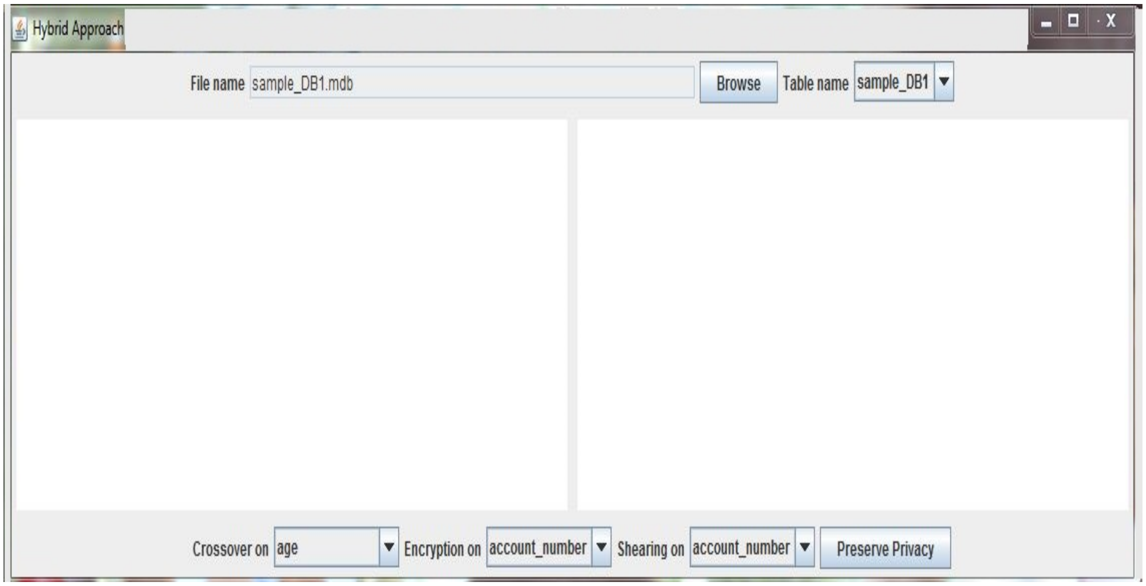


Figure-4. User interface.

As it can be seen, the user can choose the path of the specified database by clicking on the Browse button. All the tables present in the given database will be displayed and the user can accordingly choose the required table. On selecting the table, all the fields in it will become available to help the user choose on which specified field he/she would like to apply the specific technique to.



File name sample_DB1.mdb

Table name sample_DB1

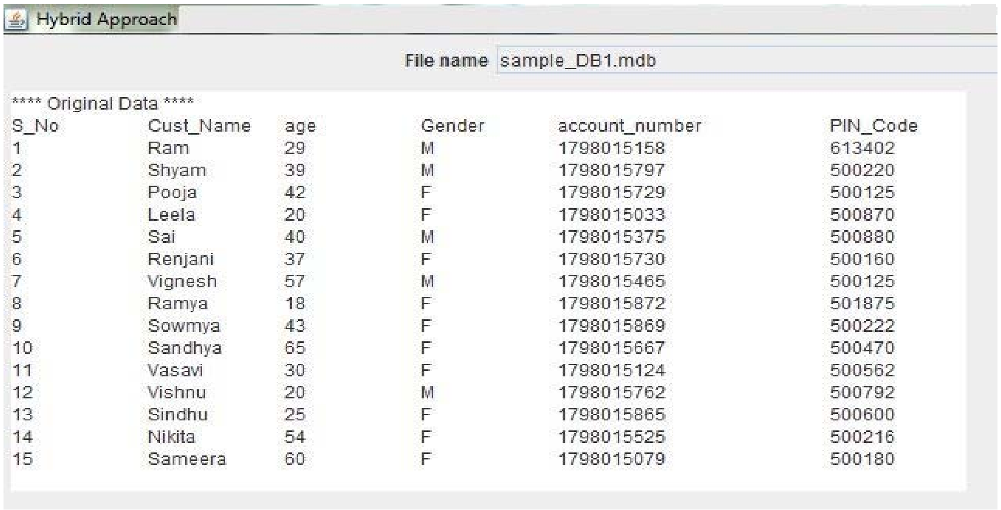
Crossover on age

Encryption on account_number

Shearing on account_number

Figure-5. On choosing the appropriate database file.

On clicking the Preserve Privacy button, the specified techniques are applied to the selected fields. The changes are accordingly made in the database. In addition, a sample of the original data and the modified data are displayed in the user interface for the user to view the differences.



File name sample_DB1.mdb

**** Original Data ****

S_No	Cust_Name	age	Gender	account_number	PIN_Code
1	Ram	29	M	1798015158	613402
2	Shyam	39	M	1798015797	500220
3	Pooja	42	F	1798015729	500125
4	Leela	20	F	1798015033	500870
5	Sai	40	M	1798015375	500880
6	Renjani	37	F	1798015730	500160
7	Vignesh	57	M	1798015465	500125
8	Ramya	18	F	1798015872	501875
9	Sowmya	43	F	1798015869	500222
10	Sandhya	65	F	1798015667	500470
11	Vasavi	30	F	1798015124	500562
12	Vishnu	20	M	1798015762	500792
13	Sindhu	25	F	1798015865	500600
14	Nikita	54	F	1798015525	500216
15	Sameera	60	F	1798015079	500180

Figure-6. Before clicking on preserve privacy button, displaying a set of original data.

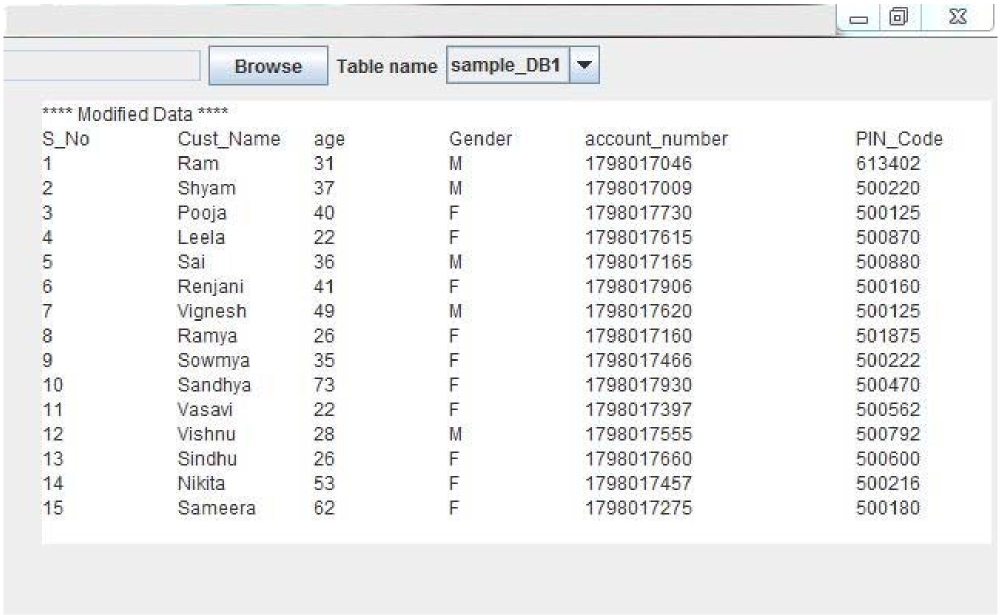


Table name sample_DB1

**** Modified Data ****

S_No	Cust_Name	age	Gender	account_number	PIN_Code
1	Ram	31	M	1798017046	613402
2	Shyam	37	M	1798017009	500220
3	Pooja	40	F	1798017730	500125
4	Leela	22	F	1798017615	500870
5	Sai	36	M	1798017165	500880
6	Renjani	41	F	1798017906	500160
7	Vignesh	49	M	1798017620	500125
8	Ramya	26	F	1798017160	501875
9	Sowmya	35	F	1798017466	500222
10	Sandhya	73	F	1798017930	500470
11	Vasavi	22	F	1798017397	500562
12	Vishnu	28	M	1798017555	500792
13	Sindhu	26	F	1798017660	500600
14	Nikita	53	F	1798017457	500216
15	Sameera	62	F	1798017275	500180

Figure-7. After clicking on preserve privacy button, displaying a set of modified data.

4. ANALYSIS

Performance Analysis is done evaluate the loss of data in applying the hybrid approach. Bias in mean (BIM) and Bias in Standard deviation (BIS) for each technique is found.

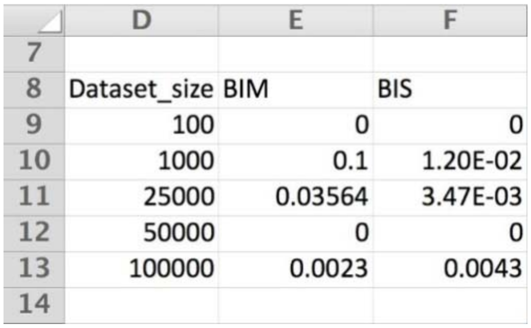
BIM = (average of perturbed data-average of original data)/ average of original data

BIS = (standard deviation of perturbed data-standard deviation of original data)/ standard deviation of original data

Lower the value of BIM and BIS better the quality of the technique

Here is the value run on sample data and the graph for those values

BIM and BIS values for Crossover:



	D	E	F
7			
8	Dataset_size	BIM	BIS
9	100	0	0
10	1000	0.1	1.20E-02
11	25000	0.03564	3.47E-03
12	50000	0	0
13	100000	0.0023	0.0043
14			

Figure-8. Table representing BIM and BIS values for Crossover on different sizes of dataset.

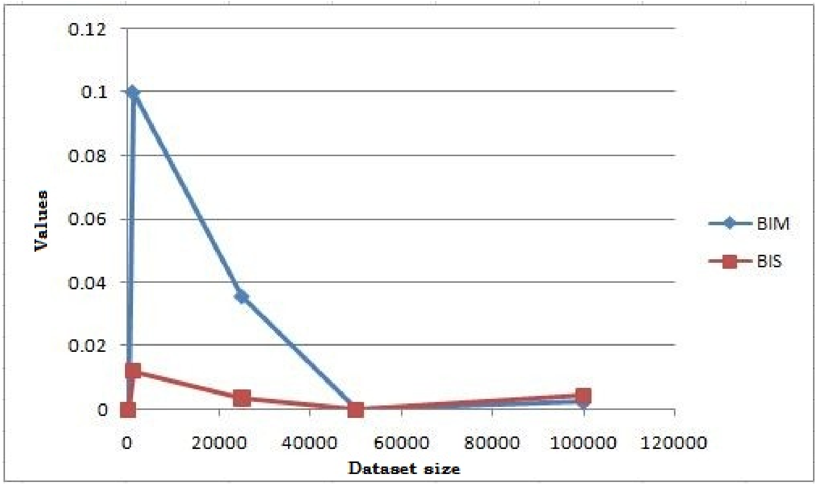
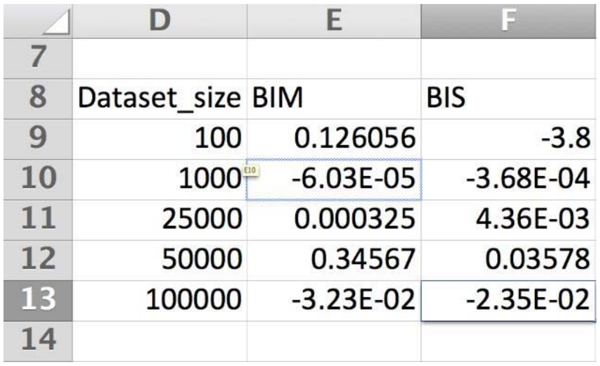


Figure-9. Graph plotted against BIM and BIS values vs different sizes of dataset for Crossover.

The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

BIM and BIS values for encryption



	D	E	F
7			
8	Dataset_size	BIM	BIS
9	100	0.126056	-3.8
10	1000	-6.03E-05	-3.68E-04
11	25000	0.000325	4.36E-03
12	50000	0.34567	0.03578
13	100000	-3.23E-02	-2.35E-02
14			

Figure-10. Table representing BIM and BIS values for encryption on different sizes of dataset.

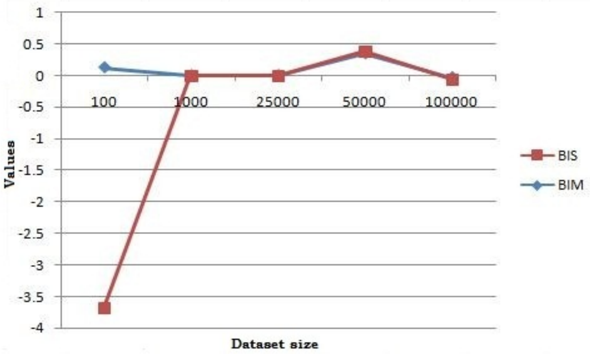
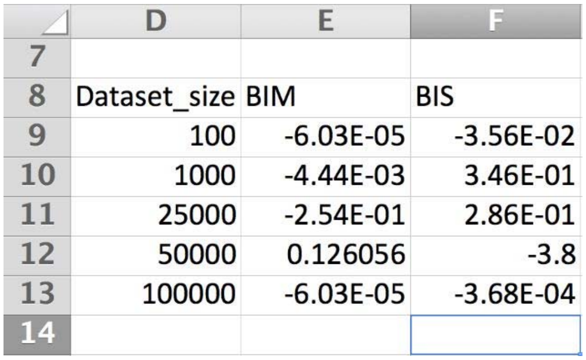


Figure-11. Graph plotted against BIM and BIS values vs different sizes of dataset for Encryption.

The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

BIM and BIS values for Shearing:



	D	E	F
7			
8	Dataset_size	BIM	BIS
9	100	-6.03E-05	-3.56E-02
10	1000	-4.44E-03	3.46E-01
11	25000	-2.54E-01	2.86E-01
12	50000	0.126056	-3.8
13	100000	-6.03E-05	-3.68E-04
14			

Figure-12. Table representing BIM and BIS values for Shearing on different sizes of dataset.

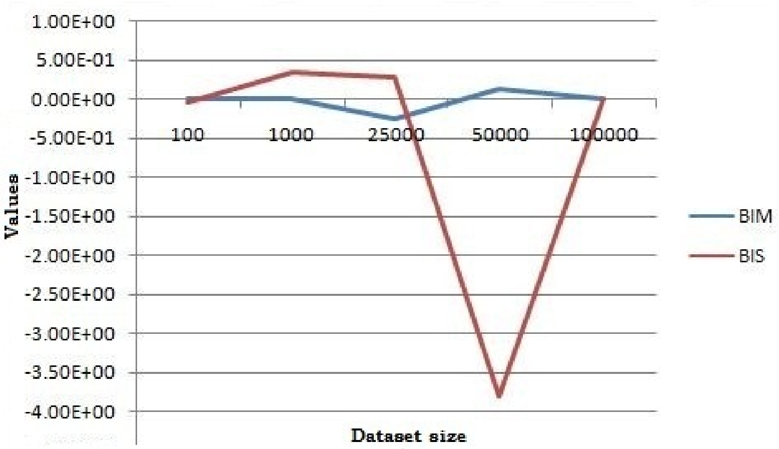


Figure-13. Graph plotted against BIM and BIS values vs different sizes of dataset for Shearing.

The values of BIM and BIS keep fluctuating because it depends on the values of the attributes present in the database. In spite of this, the value is low thus ensuring negligible data loss and high level of privacy.

5. CONCLUSIONS

This paper explains the hybrid approach for privacy preservation during the process of data publishing. Three major techniques are applied, crossover, double pass matrix based encryption and shearing. In crossover, two attribute values (field chosen by the user to which crossover is to be applied) are selected in a sequential manner and then converted into binary values. The bits are paired up. A number is generated at random and the bits in that particular position are swapped. In addition, mutation is applied if needed. During double pass matrix based encryption, the significant digits of the chosen attribute are extracted, a corresponding matrix is generated in a sequential manner, and the data is encrypted by finding its analogous position in the matrix. This process is repeated twice. In shearing, the significant digits are extracted and sheared (similar to the shearing process of graphics) by multiplying it with a random number. An average of the sheared and the original data is taken and the original data is replaced with the modified data. Further modification that can be implemented for the proposed system in the future is that the technique can be altered to include all data types.

REFERENCES

A.H.M. Sarowar Sattar, Jiuyong Li, Xiaofeng Ding, Jixue Liu, Millist Vincent. 2013. A general framework for privacy preserving data publishing. Journal homepage: www.elsevier.com/locate/knosys, Accepted 23 September 2013, Available online.

M. Naga lakshmi and K Sandhya Rani. 2013. Privacy preserving hybrid data transformation based on svd. International Journal of Advanced Research in Computer and Communication Engineering. 2(8).

Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi and Ratna Babu Chekka. 2013. A Hybrid Algorithm for Privacy Preserving in Data Mining. I.J. Intelligent Systems and Applications, 2013, 08, 47-53 Published Online July 2013 in MECS.

Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham. 2007. The applicability of the perturbation based privacy preserving data mining for real-world data. Published Online.

Nissim Matatov, Lior Rokach, Oded Maimon. 2010. Privacy-preserving data mining: A feature set partitioning approach. Accepted on.

Md Zahidul Islam, Ljiljana Brankovic. 2011. Privacy preserving data mining: A noise addition framework using a novel clustering technique. Available Online.

Weijia Yang, Sanzheng Qiao. 2009. A novel anonymization algorithm: Privacy protection and knowledge preservation.

Amar Paul Singh, Ms Dhanshri Parihar. 2013. A Review of Privacy Preserving Data Publishing Technique.

Geetha Mary A, N.Ch.S.N. Iyengar. 2011. Non-Additive Random Data Perturbation for Real World Data.