ISSN:0975-9646



# Randomized Additive Data Perturbation and Reconstruction Technique to Approximate Distribution of Original Information in PPDM

Dr. P.Kamakshi

*Department of Information Technology*

*Kakatiya Institute of Technology and Science*
*Warangal, India*

***Abstract*— Data mining or knowledge discovery is the process of analysing data from different perspectives and summarizing it into useful information. The summarized information can be utilized to increase revenue, cut costs, or both in many organizations. One of the novel research areas in data mining is to develop the techniques which can have the feature to comprise data mining as well as privacy preservation. Various techniques of privacy preservation are used to guard the privacy of an individual. Additive Randomization is one of privacy preservation technique which protects the privacy by modifying the original sensitive data and releasing the modified information to the user for analysis purpose. The original sensitive information is modified in such a manner that the statistical properties of the database do not change. In this paper we focus on additive randomization process and reconstruction procedure to approximate the original information from perturbed data. We also give the analysis which indicates the limit till which the additive perturbation can be performed without any problem and also reveals satisfactory modified data for analysis purpose or various data mining applications.**

***Keywords*— Data mining, privacy preservation, data modification, reconstruction, threshold limit.**

## I. INTRODUCTION

The powerful tool data mining can investigate [9] and extract previously unknown patterns from huge databases. The process of data mining requires a large amount of data to be collected and stored into a centralized database. Today many organizations are remarkably dependent on data mining results to provide enhanced service, accomplishing better profit and better decision-making. With rapid growth in technology and internetworking various organizations have the ability to collect and store huge amount of information for analysis purpose. Huge amount of data is collected from various sources like government offices, healthcare system, insurance companies etc. Various business organizations collect data about the consumers for marketing purposes and improving business strategies, medical organizations collect medical records for better treatment and medical research.

Huge volumes of data collected in this manner also include sensitive data about individuals. It is obvious that if a data mining operation is performed on such databases, the extracted consolidated report or knowledge may consists of patterns and correlations that are hidden in the data but it also reveals the information which is considered to [7] be private. In many data mining applications that deal health care, security, financial and other types of sensitive data privacy preservation is an important issue. The actual anxiety of people is that their private information should not be misused behind the scenes without their knowledge. The real threat is that once information is unrestricted, it will be impractical to stop misuse.

The simplest solution to this problem is to completely hide the sensitive data or not to include such sensitive data in the databases. Still, privacy can be violated when data mining techniques uses the identifiers which themselves are not very sensitive, but are used to identify an individual by connecting personal identifiers such as addresses, names etc.

But this simple and straightforward solution is not ideal and accurate for many applications like banking, scientific and medical research, DNA research etc., because today organizations realized and wish to conduct a joint research on their databases because combining their data will definitely provide better results and mutual benefit to the organizations. In this situation organizations want to share the data but restrict themselves due to privacy concern about their clients information. In such circumstances it is not only essential to shield private and sensitive information but it is also essential to facilitate the use of database for investigation or for other purposes. Privacy preserving data mining is a unique data mining technique which has emerged to shield the privacy of sensitive data and also give valid data mining outcome.

## II. PREVIOUS WORK

Recently the application of data mining is increased in various domains like business, academia, communication, bioinformatics and medicine. The data mining results not only gives the valuable information hidden in these databases, but sometimes also reveals private information about individuals. The difficulty is that data mining process extracts or evaluates the individual data which is considered as private by means of linking different attributes. The true problem is not data mining, but the way data mining is done. PPDM is an emerging technique in data mining where privacy and data mining can coexist. It gives the summarized results without any loss of privacy through data mining process.

## III. PRIVACY PRESERVING DATA MINING METHODS

For the past few years, several approaches have been proposed in the context of [1] privacy reserving data mining. These techniques can be classified based on the different protection methods used, such as Data modification methods, Cryptographic methods. Fig-1 shows the classification of different privacy preserving data mining methods.

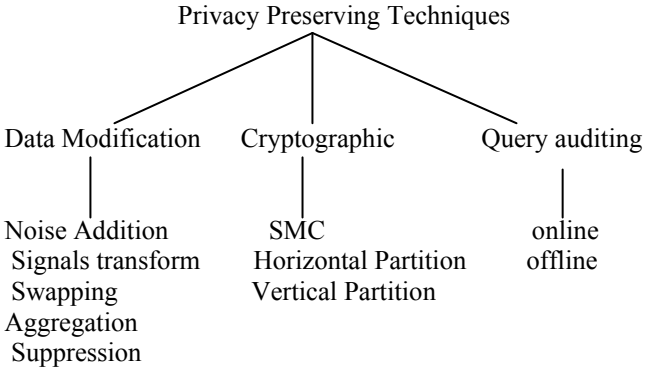



Fig 1 Classification of privacy preserving methods

*A .Data modification* technique modifies the data before releasing it to the users. Data is modified in such a way that the privacy is preserved in the released data set.

*B .Cryptographic methods* utilizes encryption and decryption the sensitive data and also allow the data for [5] data mining tasks. Protocol such as secured multiparty computation (SMC) does not disclose any new information other than the final result of the computation to a client. .

*C. Query auditing* methods privacy is preserved by modifying or restricting the results of a query.

## III. OVERVIEW OF ADDITIVE DATA MODIFICATION TECHNIQUE

Perturbation techniques preserve the privacy of individual sensitive data by altering the original data with some known distribution of noise. Here the users are provided access only to the modified values instead of original values. The main usage of perturbation techniques comes where there is a need to provide the data to a third party for data mining to retrieve hidden patterns. In randomization approach [2] the privacy of the data is obtained by perturbing it with randomization algorithms and submitting the randomized version, thus hiding the data and guaranteeing protection against the reconstruction of the data. In this scheme, a random number is added to the value of a sensitive attribute. For example, if X is the value of a sensitive attribute than, Xi+r will appear in the database, where r is a random value drawn from some distribution. This method is known as additive data perturbation. Most commonly used distributions is Gaussian distribution with mean equal to zero and .standard deviation s. The algorithm is so chosen that aggregate properties of the data can be recovered with sufficient precision while individual entries are significantly distorted. The server has a complete and precise database with information from its clients, and it has to make a version of this database public for others to work with. The method of randomization can be described as follows. Consider a set of data records denoted by X = {x1 . . xN}. For record xi Є X, we add a noise component which is drawn from the probability distribution $f_y(y)$. These noise components are drawn independently, and are denoted y1 . . . $y_N$. Thus, the new set of distorted records are denoted by x1+y1 . . . xN +yN. We denote this new set of records by z1 . .. zN. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data.

One key advantage of the randomization method [ 6] is that it is relatively simple, and does not require knowledge of the distribution of other records in the data.

## V. RECONSTRUCTION PROCEDURE

Given a cumulative distribution $F_y$ and the realizations of n random samples$X_1+Y_1$, $X_2+Y_2$,….,Xn+Yn . $F_x$ is estimated.

Let the value of $X_i + Y_i$ be $w_i$ (= $x_i + y_i$). Baye's rule is used to estimate the posterior distribution function $F^1_{x1}$ (given that $X_1+Y_1 = w_1$ ) for X1, assuming we know the[10] density functions fx and fy for X and Y respectively. **Baye's rule** is given as follows

An event A corresponds to a number of exhaustive events B1,B2,B3…….,Bn.If $P(B_i)$ and $P(A/B_i)$ are given, then

$$P(B_i/A)=(P(B_i.P(A/B_i))/(\sum P(B_i.P(A/B_i))$$

First, from the given data minimum and maximum values are obtained . Then data is distributed 10 intervals based on the minimum and maximum values. Then, posterior distribution function is obtained for each interval using Baye's rule.

From the algorithm we can observe that for each iteration the value of each interval is improved and it gets [8] closer to it. We have to stop the iteration, whenever the two successive distributions have the difference less than 1%.

## VI. PROPOSED ARCHITECTURE

With the development of technology and networking organizations collect and store huge amount of data .Such huge volume of data is considered as very important asset for an organization. Most of The companies use data mining tools to extract the unknown [4] pattern from such data. The data owners use such interesting patterns for analysis and decision making process**.** In today's competitive world the companies realized that growth of the organization is not possible by only an individual but it is possible only by means of sharing information and collaborating with other companies .But the companies restrict themselves because of the limitation on sharing of private data. The additive perturbation technique protects the privacy of the sensitive information by adding small of noise to the original data and submitting the perturbed data to the outside world for analysis purpose.

The process diagram consists of following steps:

a) Client submits the query to the data miner. The data miner interact with other parties having their own databases and are under the collaboration with each other but don't trust each other .
b) The Data miner submit the query to the database owner.
c) The database owner analyses the query and identifies whether the required attributes values are sensitive or not. If the required attribute values are not sensitive the original data values for that particular attributes are revealed to the data miner without any modification.
d) If required attributes are sensitive then following operations are performed :
   i) The sensitive attribute values are modified using randomized additive data modification technique with different values of standard deviations.
   ii) All the modified valued with different standard deviations are reconstructed. Only the modified values below the threshold limit are released to data miner for analysis purpose.

The main reason of revealing only the modified values below the threshold limit is that because these values provides approximate distribution of original values and statistical properties of data is maintained. If we reveal the modified values whose standard deviation is more than threshold limit then the problem is that the original distribution of data cannot be achieved. the modified values may exceed the lower and upper limit of the data, which in turn will violate the statistical properties of database and the released data will not be suitable for analysis or data mining purpose ,though it preserves the privacy of original sensitive information.

The interaction between the client, data miner and database owner is shown in Fig. 2

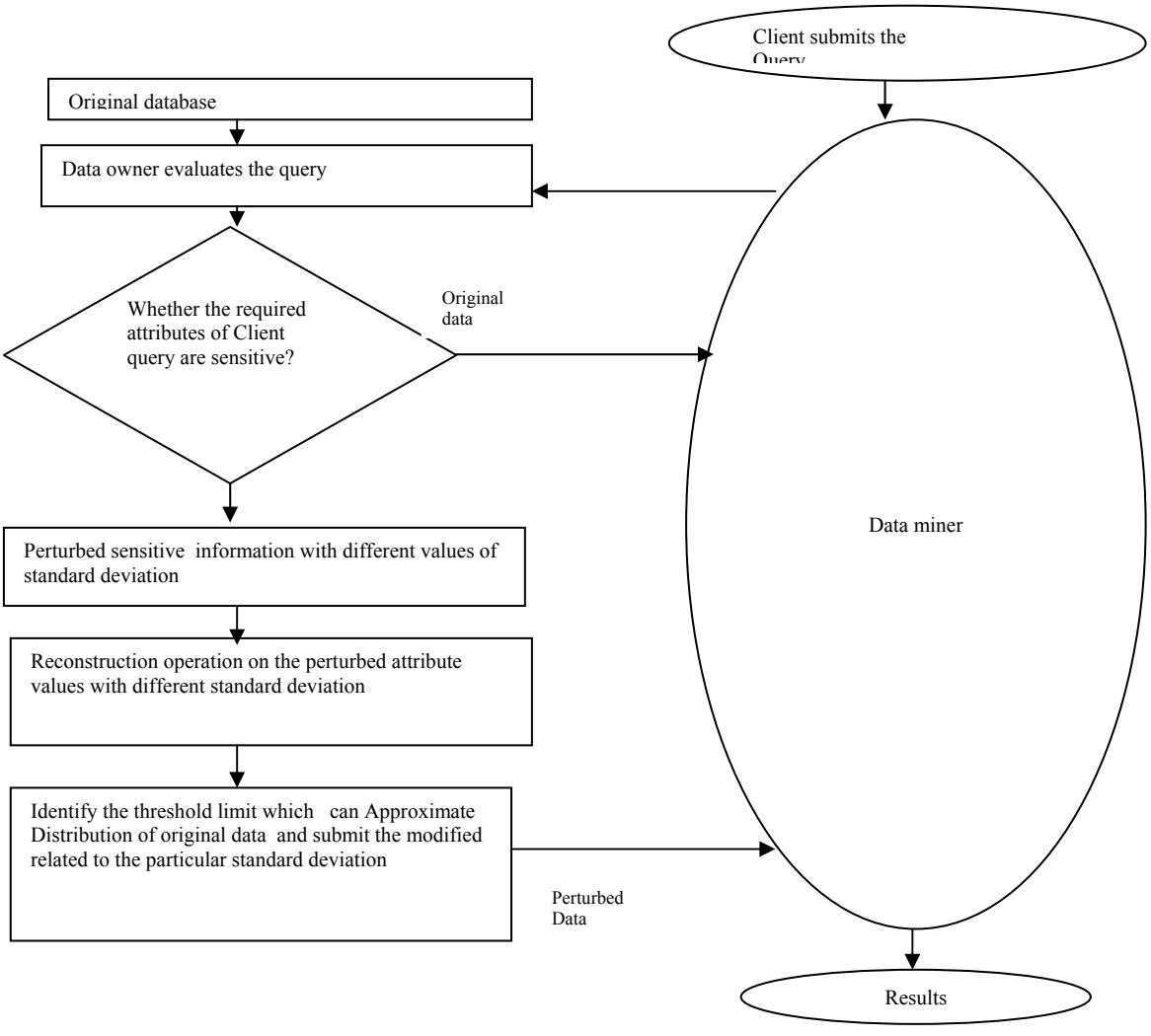



Fig.2.Additive perturbation and reconstruction procedure in PPDM

Healthcare database is taken for evaluation purpose. There are number of attributes are there in this database, but we considered the attribute length as sensitive attribute for analysis purpose. Data modification is performed with various standard deviations and again reconstruction procedure is also applied. We analysed perturbed as well as reconstructed values and the perturbed values are released for the standard deviation of which the original distribution of data values can be approximated. We considered the SD of 2,3,4,5,6,7,8,9 and 10.It is found that with SD of 4 one can release the perturbed sensitive information because the distribution of original data can be approximated. Beyond SD=4, the variation between original, perturbed and approximated value is very large and such modified information cannot be released for analysis purpose because it is causing change in the original statistical properties of database.

## VII. RESULTS

We performed analysis on healthcare database with Standard deviation =2 to standard deviation =10 considering attribute length as sensitive attribute. Fig 3 shows the sample of original database. Fig 4 and Fig.5 show the tables of original perturbed and reconstructed values. For analysis purpose we have taken the SD=2,4 and 10. Simultaneously we performed reconstruction operation on the perturbed data. Fig 6, Fig. 7 and Fig.8 shows the graphs related to SD= 2,4,10 which clearly shows the variation between original, perturbed and reconstructed values. After analysis it is observed that:

a) With standard deviation of value 2, the variation between original, perturbed and reconstructed values very less. Hence privacy preservation cannot be achieved effectively.
b) With SD = 4, the variation between original, perturbed and reconstructed values are significant and also valid for giving valid data mining results.
c) With SD =10 there is large deviation between the original, perturbed and reconstructed distribution. The modified values though protects privacy , are not valid for data mining operation because the values does not maintain the statistical properties of original database.

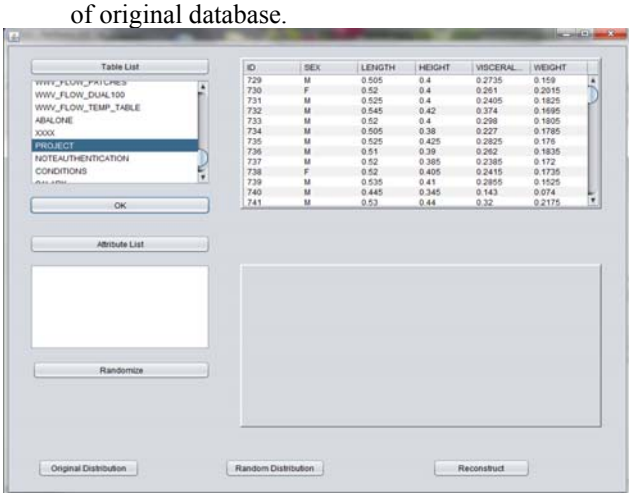



Fig 3. Sample of original database

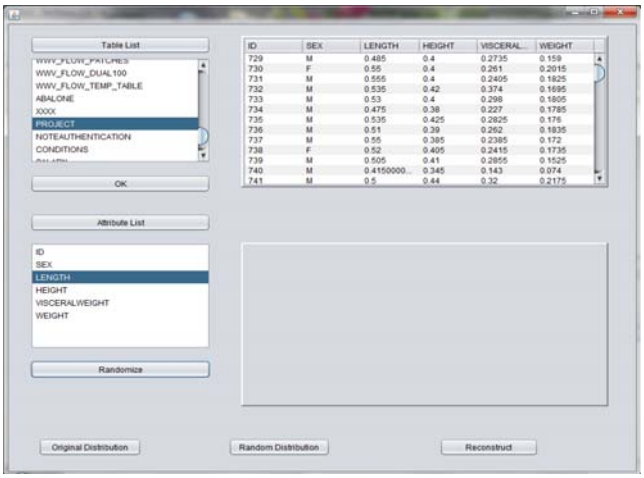



Fig4. Perturbed data

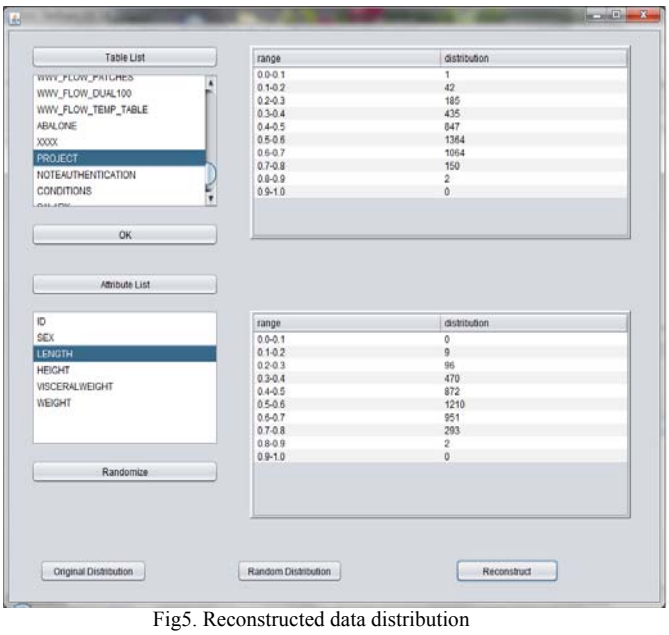



Fig5. Reconstructed data distribution

Comparison of data distributions with different standard deviations are shown below.

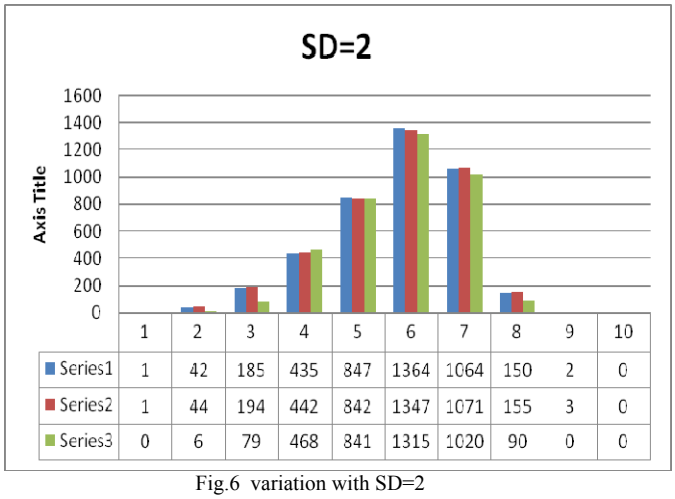
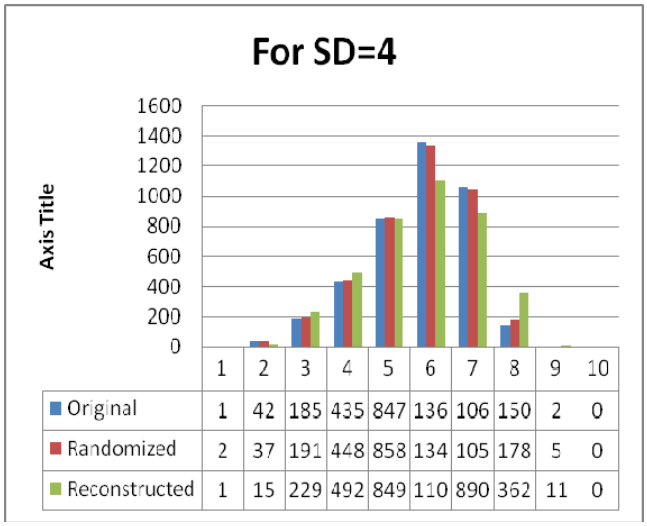
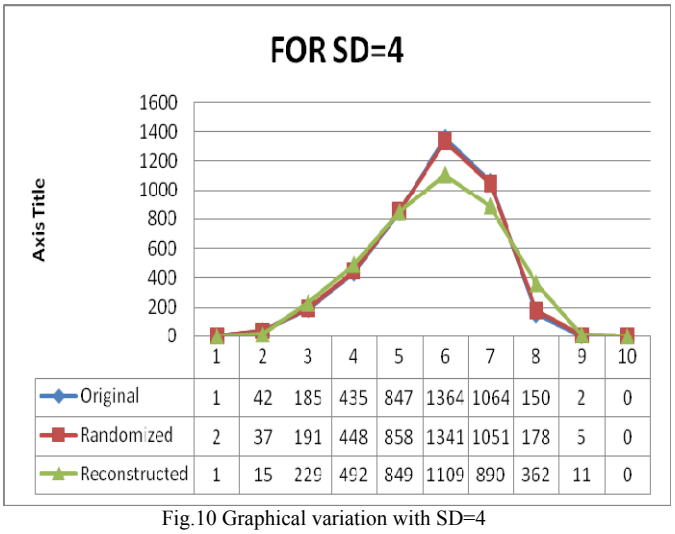



Fig.6 variation with SD=2



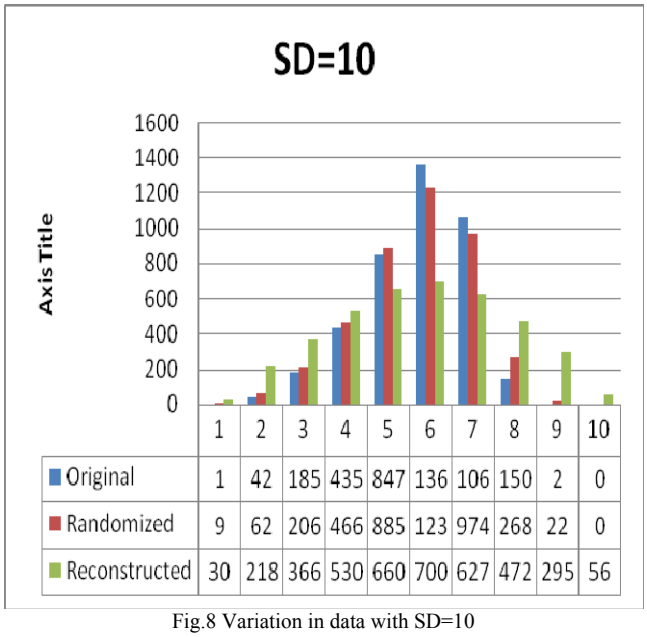

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 42 | 185 | 435 | 847 | 136 | 106 | 150 | 2 | 0 |
| Randomized | 2 | 37 | 191 | 448 | 858 | 134 | 105 | 178 | 5 | 0 |
| Reconstructed | 1 | 15 | 229 | 492 | 849 | 110 | 890 | 362 | 11 | 0 |

Fig.7 variation with SD=4




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 42 | 185 | 435 | 847 | 1364 | 1064 | 150 | 2 | 0 |
| Randomized | 2 | 37 | 191 | 448 | 858 | 1341 | 1051 | 178 | 5 | 0 |
| Reconstructed | 1 | 15 | 229 | 492 | 849 | 1109 | 890 | 362 | 11 | 0 |

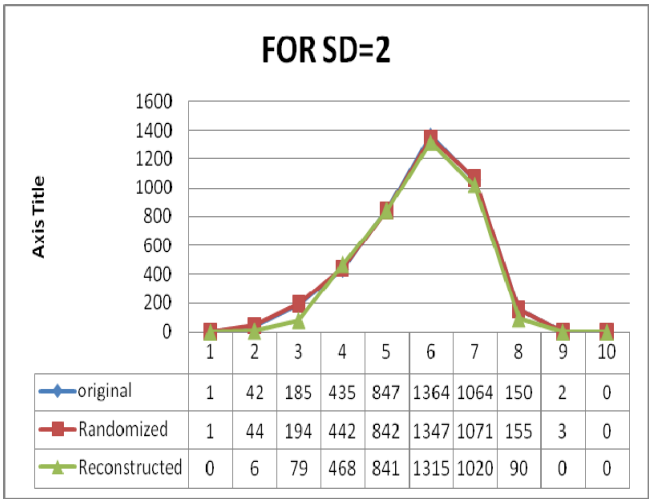
Fig.10 Graphical variation with SD=4




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 42 | 185 | 435 | 847 | 136 | 106 | 150 | 2 | 0 |
| Randomized | 9 | 62 | 206 | 466 | 885 | 123 | 974 | 268 | 22 | 0 |
| Reconstructed | 30 | 218 | 366 | 530 | 660 | 700 | 627 | 472 | 295 | 56 |

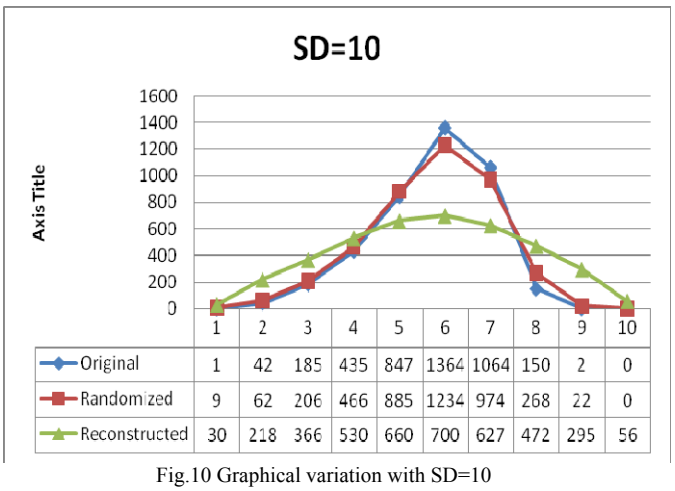
Fig.8 Variation in data with SD=10

Fig.9, Fig10 and Fig.11shows the graphical representation indicating the variation between original, perturbed and reconstructed values.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| original | 1 | 42 | 185 | 435 | 847 | 1364 | 1064 | 150 | 2 | 0 |
| Randomized | 1 | 44 | 194 | 442 | 842 | 1347 | 1071 | 155 | 3 | 0 |
| Reconstructed | 0 | 6 | 79 | 468 | 841 | 1315 | 1020 | 90 | 0 | 0 |

Fig. 9 Graphical variation with SD=2




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1 | 42 | 185 | 435 | 847 | 1364 | 1064 | 150 | 2 | 0 |
| Randomized | 9 | 62 | 206 | 466 | 885 | 1234 | 974 | 268 | 22 | 0 |
| Reconstructed | 30 | 218 | 366 | 530 | 660 | 700 | 627 | 472 | 295 | 56 |

Fig.10 Graphical variation with SD=10

## VIII. CONCLUSION

In today's world sharing of information is essential as well as beneficial for the growth of an organization .But many companies restrict themselves by sharing of information and data mining results because of privacy and security concerns. This paper focus on one of PPDM technique called randomized additive data modification technique and reconstruction procedure. Sensitive data in the database will be perturbed using a randomizing function so that they cannot be estimated with sufficient precision. Randomization can be done using Gaussian or uniform perturbations .Reconstruction of the original data distribution is performed to check whether the statistical properties of original database are maintained after perturbation or not. Perturbed information with original data distribution closer to the original data distribution is only released to the outside world or data miner for analysis purpose. As we increase the standard deviation value, the reconstructed data distribution is not closer to the original data distribution, which in turn will not reveal valid data mining results.

## REFERENCES


[1] R.Agarwal and R.Srikant, "Privacy preserving data mining",In Procseedings of the 19th ACM SIGMOD conference on Management of Data ,Dallas,Texas,USA, May2000.

[2] K.Muralidhar.,R.Sarathi, "A General additive data perturbation method for data base security" journal of Management Science. ,45(10):1399-1415,2002

[3] K. Muralidhar and R. Sarathy. Data shuffling - a new masking approach for numerical data. Management Science, 52(5):658–670, May 2006.

[4] R.Agrawal, A.Evfimievski, R.Srikant, "Information sharing across private databases", *In Proc.of ACM SIGMOD*, 2003.

[5] B. Pinkas, " Cryptographic techniques for privacy preserving data mining " *SIGKDD Explorations,* 12–19, 2002

[6] D. Agrawal and C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. Proceedings of PODS, pages 247-255, 2001.

[7] M. Kantarcio˘glu, J. Jin, and C. Clifton. When do data mining results violate privacy? In Proceedings of the 10th ACM SIGKDD Conference (KDD'04), pages 599–604, Seattle, WA, August 2004

[8] C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. ACMTransactions onDatabase Systems (TODS), 10(3):395– 411, 1985.

[9] Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber 2nd edition,Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.

[10] H. Cramer . Mathematical Methods of Statistics. Princeton University Press,1946.