A Random Approach for Privacy Preserving Data Mining by Using M-Privacy

Nagendra Patel[1], Mr. Prateek Gupta[2]

[1]*Mtech student SRIST Jabalpur India*

[2]*Asst. Professor SRIST Jablpur India*

Abstract — ***Privacy preserving data mining addresses the issues related to maintaining the data privacy while publishing the data. The privacy of the exact data is maintained by publishing the perturbed copies of the data. Traditionally the data is perturbed by mixing the Gaussian noise data. The level perturbation is depends upon the trust (the more trusted a data miner can access the less perturbed copy of the data) on the party for which the data is required to generate. In practical situations the same data is perturbed for different level of trusts (for different uses) or same level of trusts, such situations arise the threat on security of data that may cause estimation of accurate data (especially when multiple copies of data mixed with Gaussian noise are available) from these copies by advance computational algorithm like Linear Least Squares Error (LLSE). This paper presents a nonlinear system based chaotic signal generator for perturbation of original data the complex characteristics of the chaotic signal makes it difficult for the estimators to estimate original data because the estimators works on the known noise probability distribution function (PDF). The simulation result shows that the proposed algorithm gives higher estimation error when compared with traditional algorithms.***

Keyword — ***Multilevel Trust, Data Security, Signal Estimation, Chaotic System.***

I. INTRODUCTION

The data preserving mining is used where the data owners need to export/publish/outsource the private data (such as network traffic, financial and health care). Data perturbation techniques are the most popular and simple models for privacy sensitive data in Privacy preserving data mining applications (PPDM). It is low computational cost and simplicity makes it convenient for such applications. In general the perturbation procedure is described as when the data owner needs to provide their sensitive data to service providers that are not within the trust boundary it randomly changes in the private data in a way that removes the sensitive information while preserves the particular data property that is critical for analysis the data models. In present many perturbation techniques are available, among which the most typical ones are randomization approach and condensation approach.

II. RELATED WORK

The data mining is one of the very popular data intensive tasks, Privacy preserving data mining (PPDM) for the outsourced data has become an important enabling technology for utilizing the public computing resources many proposals are published in the related fields some of them are discussed in this section. Kun Liu et al [2] explore Independent Component Analysis as a possible tool for breaching privacy in deterministic multiplicative perturbation-based models such as random orthogonal transformation and random rotation and proposed an approximate random projection-based technique to improve the level of privacy protection while still preserving certain statistical characteristics of the data. Yaping Li et al [5] proposed Multi-level Trust in Privacy Preserving Data Mining in which they provide the solution against diversity attacks which applied for data miners who have access to an arbitrary collection of the perturbed copies can be jointly used for reconstructing the pure data more accurately than the best effort using any individual copy in the collection. Geometric data perturbation is proposed by Keke Chen et al [6] they argue that selectively preserving the task/model specific information in perturbation will help achieve better privacy guarantee and better data utility. This type of such information is the multidimensional geometric information, which is certainly utilized by many data-mining architectural models. To preserve that information in data perturbation, they propose the Geometric Data Perturbation (GDP) method. Michael M. Groat et al [7] present introduce multidimensional adjustment, a method that reduces the increase of error associated with earlier work for privacy-preserving participatory sensing scheme for multidimensional data which uses negative surveys. Glenn M. Fung et al [8] proposed linear programming based approximation, which is public but doesn't reveal the privately held data matrix A, has accuracy comparable to that of an ordinary kernel approximation based on a publicly disclosed data matrix A. Privacy preserving technique using Non-metric Multidimensional Scaling is proposed in [9], which not only preserves privacy but also maintains data utility for Support Vector Machine (SVM) classification. Antoine Boutet et al [11] propose a mechanism to preserve privacy while leveraging user profiles in distributed recommender systems. Their approach relies on (i) an original obfuscation mechanism hiding the exact profiles of users without significantly decreasing their utility, as well as (ii) a randomized dissemination algorithm ensuring differential privacy during the dissemination process. An overview of project PREDICT (Privacy and Security Enhancing Dynamic Information Collection and monitoring) is presented by Li Xiong et al [12]. The overall aim of the project is to develop a framework with algorithms and mechanisms for privacy and security enhanced dynamic data accumulation, aggregation, and analysis with feedback loops.

III. RANDOM PERTURBATION TECHNIQUE

For the sake of perfection, now briefly review the random data perturbation technique indicated in [13][16] for hiding the original data (protection against reconstruction of original data) while static being able to estimate the prime distribution.

A) Data Perturbation Technique

The random data perturbation technique attempts to preserve data privacy by modifying values of the sensitive attributes by using a randomize process [2].The authors find out two presumable approaches, Value class membership and Value distortion approach, the data owner returns a value ui + v, where ui is the original data and v is a random noise generated by chaos signal generator and generated from the particular distribution. Mostly we used uniform distribution over an interval [-α, α] and Gaussian distribution with mean μ = and standard deviation . The n pure data values u1, u2, u3, u4......ui are viewed as receipt of n independent and identically distributed random variables Ui, i=1, 2, 3, 4, n each are same distribution as that the random variable U. In a manner to perturb the data as n independent samples v1, v2, v3, v4......vn are drawn form a distribution. The data owner provides a perturbed values ui + vi and a cumulative distribution function Fv(r) of V. The data reconstruction problem is estimate the noise distribution Fu(x) of the genuine data, from the noisy data.

B) Estimate the distribution function from Perturb dataset

In this paper authors **[**13][16] counsel the under mentioned technique to estimate the distribution FU(u) of U, disposed independent samples ui + vi and Fu(V). By using bayes rule, the posticous distribution function FU (u) of U. Given that U + V=w can be under mentioned

$$F'_U(u) = \frac{\int_{-\infty}^{u} f_V(w-z) f_U(z)dz}{\int_{-\infty}^{\infty} f_V(w-z) f_U(z)dz},$$

Differentiation of F'U (u) with respect to u yields the density function

$$f'_U(u) = \frac{f_V(w-u) f_U(u)}{\int_{-\infty}^{\infty} f_V(w-z) f_U(z)dz},$$

Where $F_u(.)\ and\ F_v(.)$ indicate the probability density function (PDF) of $U\ and\ V$ respectively. If we take a data independent sample $u_i + v_i, i = 1,2,3 \dots n$ in that behalf posticous distribution can be acquired by averaging

$$f'_U(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_V(w_i-u) f_U(u)}{\int_{-\infty}^{\infty} f_V(w_i-z) f_U(z)dz}. \qquad (1)$$

For adequately most samples n we expect the density function which is indicated above that are close to the real density function FU (u). In practically the true density function FU (u) is unaware, we require to changes in right hand side in equation 1. The counsel a stepwise process where each step j= 1,2,3,4......the posticous density function $f_U{}^{j-1}($ are estimated at step j-1 is used in equation 1.for initialization of the iteration we use the uniform density function for speedup of a computations, the authors also confer approximations of the beyond procedure.

IV. CHAOTIC SYSTEM

Chaos is the science of daze, of the nonlinear and the unpredictable. It teaches us to hope the unexpected. While most conventional science deals with presumedly predictable phenomena like gravity, electricity and chemical reactions, Chaos theory compromise with nonlinear things that are effectively impossible to predict or control, like discomposure, weather, the stock market, our brain conditions, and so on. These phenomena are often described by mathematics, which captures the infinite complexity of the nature. Chaotic system completely dependent on initial states a more rigorous way to express this is that small changes in the initial conditions lead to drastic changes in the results.

A) Chaotic Behaviour of Electric Circuit

Electronic circuits can generally be linear or nonlinear. As no complete linearity exists in the real world, all circuits are actually nonlinear. Their analysis is usually mathematically difficult as it is linked to solving non-linear differential equations.

Different chaotic circuits have been mentioned in numerous scientific articles. These are simple RLC-circuits, various oscillators, flip-flops, capacitive-trigger circuits, digital filters, power supplies and power circuits, adaptive filters, converters.

Among the chaotic circuits the most established one being an object of numerous scientific activities (Chua et al, 1993), is the Chua's oscillator. Kennedy asserts (Kennedy, 1993a, 1993b) that the Chua's oscillator is the only physical system for which the presence of chaos has been established experimentally, confirmed numerically (with computer simulations) and proven mathematically (Chua et al, 1986) [14].

The behaviour of chaotic circuits is orderly disordered. Experiments show that in specific conditions (chosen parameters, initial conditions, input signals etc.) almost all electric and electronic circuits behave chaotically. Chaotic circuits and other kinds of chaotic systems have certain common characteristics like: high sensitivity to initial conditions, positive Lyapunov exponents, bifurcations, fractals, chaotic attractors, etc.When using this kind of systems in cryptography, these characteristics hence transferred into cryptosystems [14].

B) Colpitts Oscillator as Chaotic System

1) Circuit Model

We consider the classical configuration of the Colpitts oscillator containing a bipolar junction transistor (BJT) as the gain element and a resonant network consisting of an inductor and a pair of capacitors, as illustrated in Fig. 1. Note that the bias is provided by the current source characterized by a Norton-equivalent conductance According to the qualitative theory in nonlinear dynamics; we select a minimal model for the circuit. The idea here is to consider as simple a circuit model as possible which maintains the essential features exhibited by the real collpitts oscillator [15].

2) Modelling

For the analysis of colpitts oscillator as chaotic system it is necessary to consider nonlinearities in the behaviour of transistor. In the system it is maintained by characterizing the B-E junction of transistor as follows

a) We model the V–I characteristic of with an exponential function are namely

$$I_E = Is\left[\exp\left(\frac{V_{BE}}{VT}\right) - 1\right] \approx I_s\left[\exp\left(\frac{V_{BE}}{VT}\right)\right], \quad \text{if } V_{BE} \gg V_T \qquad (1)$$

Where I_S is the inverse saturation current and 25mV at room temperature, and $V_{be} = 0.7V$.

3) State Equations

The state equations for the schematic in Fig. 1 are the following:

$$\begin{aligned} C_1\frac{dV_{C_1}}{dt} &= -f(V_{C_2}) + I_L \\ C_2\frac{dV_{C_2}}{dt} &= I_L - I_0 \\ L\frac{dI_L}{dt} &= -V_{C_1} - V_{C_2} - RI_L + V_{CC} \end{aligned} \qquad (2)$$

Where is the driving-point characteristic of the nonlinear resistor. This characteristic can be expressed in the form $I_E = f(V_{C_2}) = f(-V_{BE})$ and, in particular, from (1) it follows that

$$f(V_{C_2}) = I_S \exp\left(-\frac{V_{C_2}}{V_T}\right)$$

This circuit works as chaotic generator when the values of component is selected as, $C_1 = 15pf, C_2 = 15pf, R = 50\Omega, R_e = 150\Omega, \beta = 200, L = 100\mu H, V_{cc} = 5V$.

The waveforms of the generated signal from the above configurations are shown in figure 2 and 3.

V. PROPOSED ALGORITHM

The Proposed work can be described as follows:

Step 1: Initialization of the system by providing the information about the number of copies to be generated (N), level of trust and distribution of trust amongst the copies.

Step 2: Generate the N sets of 3 dimensional vectors using Gaussian PDF; these numbers will be used as the initial conditions for the collpitts oscillator.

Step 3: Read the Original copy of the data to be published.

Step 4: Determine the length of the dataset.

Step 5: calculate the power of the original data set.

Step 6: generate the N chaotic signal of same length using collpitts oscillator for each perturbed copy using the initial conditions generated at step 2.

Step 7: Normalized the power of each of the generated chaotic signal.

Step 8: to generate perturbed copies, add the chaotic signal with original data after multiplying it with factor F = (1 - Trust Level)

VI. SIMULATION RESULTS

For the simulation and analysis we have taken the dataset of Age, Iris, Income and Wisconsin from CENSUS [1].

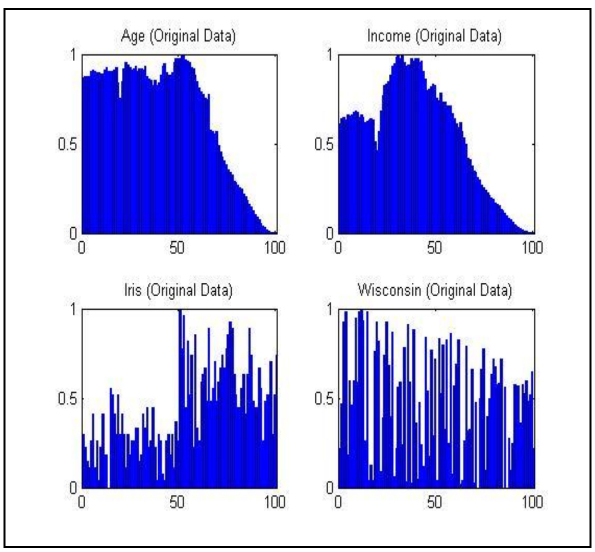

Figure 4: original plots of datasets used

Scenario 1:

Number of perturbed copies = 5

Trust Level = 0.9

Trust Distribution = Constant

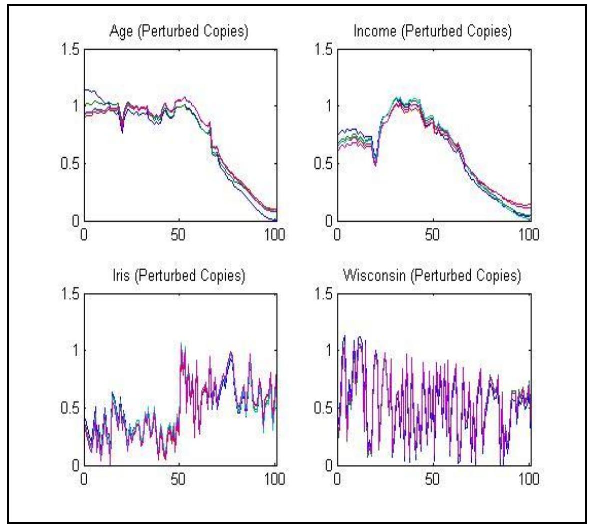

Figure 5.1: Perturbed copies of each input dataset for the current scenario.

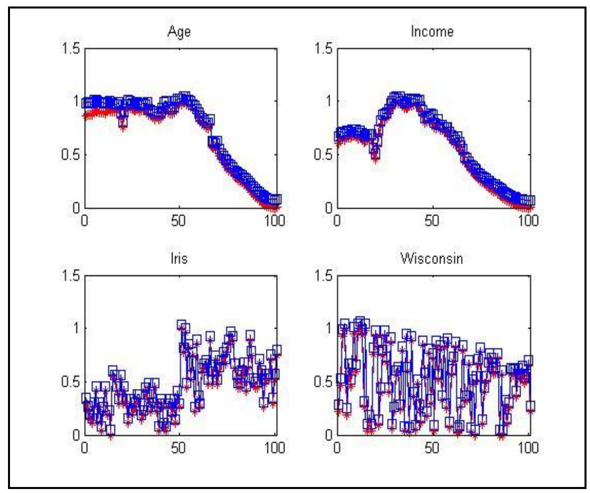

Figure 5.2: plot of estimated data using LLSE (in blue) for the original data.

Table 1: MSE comparison for different techniques:

Dataset	Standard	Previous[5]	Proposed
Age	0.0020	0.0024	0.0076
Income	0.0018	0.0023	0.0079
Iris	0.0023	0.0023	0.0079
Wisconsin	0.0023	0.0023	0.0078

Scenario 2:
Number of perturbed copies = 5
Trust Level = 0.9

Trust Distribution = Random

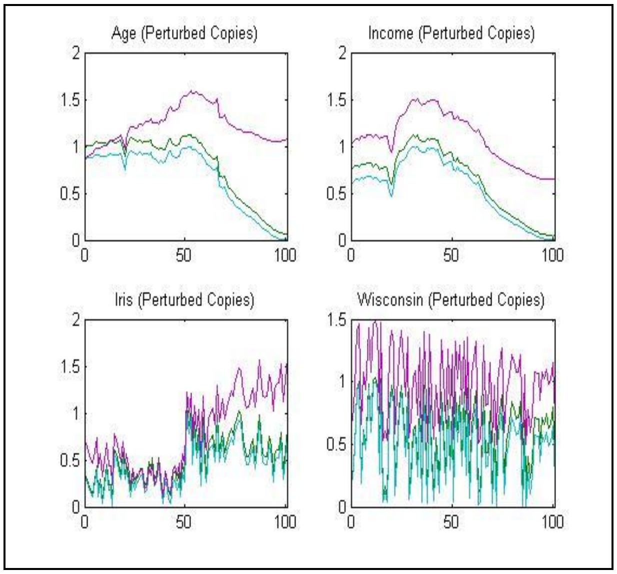

Figure 6.1: Perturbed copies of each input dataset for the current scenario.

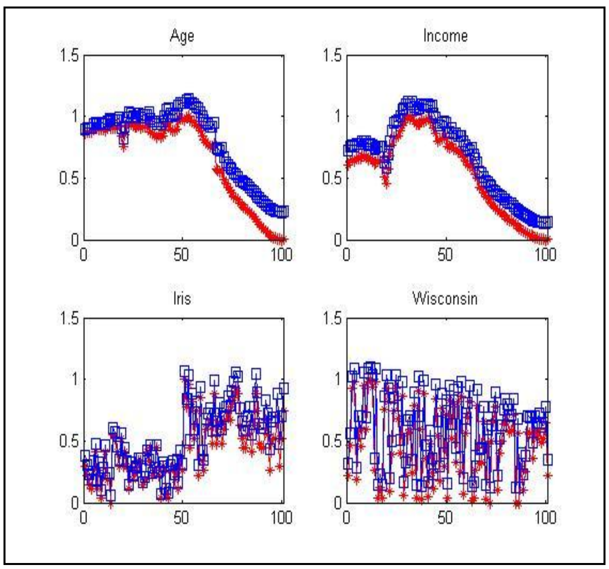

Figure 6.2: plot of estimated data using LLSE (in blue) for the original data.

Table 2: MSE comparison for different techniques

Dataset	Standard	Previous[5]	Proposed
Age	0.00506	0.0099	0.0389
Income	0.00402	0.0100	0.0403
Iris	0.00476	0.0099	0.0398
Wisconsin	0.00555	0.0100	0.0405

VII. CONCLUSION

As the importance of preserving privacy in data mining already discussed in section 1. Randomized techniques perform an important role in this paper. In this paper discuss the estimation problems that these techniques maintain the sensitive data privacy. This paper we proposed a chaotic system based data perturbation technique with multilevel trust that may find wider application in developing a new approach to preserve a sensitive data privacy in better than previous algorithm. The analysis showed that it is relatively very difficult to crack the privacy of the original data offered by the proposed method when compared with random perturbation based techniques especially when the number of published perturbed copies of same trust level increases. It also provided practical results with different types of dataset and scenarios which showed that the proposed algorithm outperforms the previous algorithm by almost three time's better security (refer to table 1 and 2).

The proposed work uses a chaotic based noise generator. While the chaotic system generates highly unpredictable noise, it contains a couple of differential equations and needed to be solved for each point of the required number of solution points. The solution of these equations may be time consuming and complex hence in future selection algorithm for best chaotic system depending upon required complexity and time can be developed.

REFRENCES

[1] A. Bednarz, N.Bean, and M.Roughon. Hiccups on the road to privacy-preserving linear and non-linear programming. In Proceedings of the 8th ACM Workshop on Privacy in the Electronics Society, pages 117-120, 2009.

[2] Kun Liu, Hillol Kargupta, Senior member, IEEE, and Jessica Ryan "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining",IEEE TRANSACTION ON KNOWLEDGE and DATA Engineering, VOL. 18, NO. 1, JANUARY 2006

[3] Charu C. Aggarwal and Philip S.Yu "A Condensation Approach to Privacy Preserving Data Mining", Advances in Database Technology - EDBT 2004 Lecture Notes in Computer Science Volume 2002, 2004, pp 183-199

[4] V. Cherkassky and F. Mulier. Learning from Data - Theory, concepts and Methods. John Wiley & Sons, New York, 1998.

[5] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang "Enabling Multi-level Trust in Privacy Preserving Data Mining", MANUSCRIPT ACCEPTED FOR PUBLICATION IN IEEE

TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2011.

[6] Keke Chen, Ling Liu ”Geometric data perturbation for privacy preserving outsourced data mining”, Knowl Inf Syst, Springer 23 November 2010.

[7] Michael M. Groat, James Horey Benjamin Edwards, Wenbo He and Stephanie Forrest “Enhancing Privacy in Participatory Sensing Applications with Multidimensional Data”, erCom, page 144-152. IEEE, (2012).

[8] Glenn M. Fung & Olvi L. Mangasarian “Privacy-Preserving Linear and Nonlinear Approximation via Linear Programming”, Optimization Methods & Software Volume 28 Issue 1, February 2013 Pages 207-216.

[9] Khaled Alotaibi and Beatriz de la Iglesia “Privacy-Preserving SVM Classification using Non-metric MDS”, SECURWARE 2013 : The Seventh International Conference on Emerging Security Information, Systems and Technologies.

[10] L. Liu, J. Wang, Z. Lin, and J. Zhang. Wavelet-based data distortion for privacy-preserving collaborative analysis. Technical Report 482-07, Department of Computer Science, University of Kentucky, Lexington, KY 40506, 2007. http://www.cs.uky.edu/ jzhang/pub/MINING/lianliu1.pdf.

[11] Antoine Boutet, Davide Frey, Arnaud Jegou, Anne Marie Kermarrec, Rachid Guerraoui ““Privacy-Preserving Distributed Collaborative Filtering”, RESEARCH REPORT N° 8253 February 2013.

[12] Li Xiong, Vaidy Sunderam, Liyue Fan, Slawomir Goryczka, Layla Pournajaf “PREDICT: Privacy and Security Enhancing Dynamic Information Collection and Monitoring”, International Conference on Computational Science, ICCS 2013 science direct.

[13] Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamurthy Sivakumar “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, Data Mining, 2003. ICDM 2003. Third IEEE International Conference on 19-22 Nov. 2003.

[14] Matej Šalamon “Chaotic Electronic Circuits in Cryptography”, University of Maribor, Faculty of Electrical Engineering and Computer Science Slovenia.

[15] Maggio, Gian Mario; de Feo, Oscar; Kennedy, Michael Peter “Nonlinear analysis of the Collpitts oscillator with applications to design”, IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications, Vol. 46, No. 9, September 1999.

[16] R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceeding of the ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, Texas, May 2000 ACM Press.