Preserving Data Mining through Data Perturbation

Mr. Swapnil Kadam, Prof. Navnath Pokale

***Abstract*— Data perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single level trust on data miners.Privacy Preserving Data Mining deals with the problem of developing accurate models about aggregated data without access to precise information or original records in individual data record. perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy of data before data are published. Previous solutions of this approach are not suitable in their tacit assumption of single-level trust on data miners. In this work, we cosidering this assumption for expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In our approach, the more trusted a data miner is, the less perturbed copy of the data it can access. Under this, a malicious data miner have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. Preventing diversity attacks is the challenge of providing MLT-PPDM services. We resolve this challenge by properly assignment perturbation across copies at different trust levels. We prove that our solution is good against diversity attacks with respect to our privacy goal. That is, for data miners who have access to an arbitrary collection of the perturbed copies, our technique prevent them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. Our method allows a data owner to generate perturbed copies of its data for as per trust levels on demand. This technique offers data owners maximum flexibility.**

***Index Terms*— Multilevel Trust, Privacy Preserving Data Mining, random perturbation.**

.

I. INTRODUCTION

Preserving privacy in Data Mining (PPDM) technique introduces uncertainty about individual values before data are published or released to third parties for data mining purposes [1], [2], [3], [4], [5], [6], [7]. In the single level trust assumption, the data owner or admin can create only single perturbed copy of its data with a fixed amount of uncertainty. So this type of assumption will limited in different applications where a data owner or admin trusts a data miners at different levels. We are going to present a two trust level scenario as a motivating example as described below. The business or a government have to do useful internal data mining which should be most trusted , but they may also want to release the data to the public, and might perturb it more. The less perturbed internal copy is received by a mining department. This mining department also has access to the more perturbed public copy. It could be desirable that mining department do not have any more authority for recreating the original data by utilizing both copies than when it has only the internal copy. Similarly, when an internal copy is discharged to the public, then obviously the public has all the power of the mining department. Although, it could be advisable if the public can't recreate the original data more accurately when it uses both copies than when it uses only the leaked internal copy. These new dimensions of Multi-level Trust (M.L.T.) poses challenges for PPDM which is based on perturbation. In contrast to the single-level trust scenario where only one perturbed copy is released, now more number of differently perturbed copies of the actual data are accessible to data miners at various trusted levels. It means the less perturbed copy can be accessed by the more trusted data miner ; it may also have access to the perturbed copies available at lower trust levels. Generally, the data miner could access multiple perturbed copies through various other means, e.g., accidental leakage or colluding with others. By utilizing diversity across differently perturbed copies, the data miner may be able to produce a more accurate reconstruction of the original data than what is allowed by the data owner. We refer to this attack as a diversity attack. It includes the colluding attack scenario where adversaries combine their copies to mount an attack; it also includes the scenario where an adversary utilizes public information to perform the attack on its own. Preventing diversity attacks is the key challenge in solving the MLT-PPDM problem .In this paper, we address this challenge in enabling MLT-PPDM services. In particular, we focus on the additive perturbation approach where random Gaussian noise is added to the original data with arbitrary distribution, and provide a systematic solution. Through a one-to-one mapping, our solution allows a data owner to generate distinctly perturbed copies of its data according to different trust levels. Defining trust levels and determining such mappings are beyond the scope of this paper.

A. Contributions

We make the following contributions.

- We expand the scope of perturbation-based PPDM to

***Manuscript received Nov, 2015*.**

***Mr Swapnil Kadam** pursuing Master's degree in Computer Engineering at Department of Computer Engg.,BSCOR, Narhe, Pune University.9423817389.*

***Prof. Navnath B. Pokale** Curently working as a Assistant Professor at Department of Computer Engg.,BSCOR, Narhe, Pune University.*

multilevel trust, to generate differently perturbed copies of its data for different trust levels and introducing random rotation based algorithm.

- We identify a key challenge in enabling MLT-PPDM services. In MLT-PPDM, data miners may have access to multiple perturbed copies. By combining multiple perturbed copies, data miners may be able to perform diversity attacks to reconstruct the original data more accurately than what is allowed by the data owner. Defending such attacks is challenging, which we explain through a case study in Section 4.
- We address this challenge by properly assigning perturbation across copies at per trust levels. We prove that our solution is robust against diversity attacks.
- We provide solution that allows data owners to generate perturbed copies of their data as per trust levels on-demand. This property offers data owners maximum flexibility.

II. LITERATURE SURVEY

Privacy Preserving Data Mining (PPDM) was first proposed in [2] and [8] simultaneously. To address this problem ,researchers have since proposed various solutions that fall into two broad categories based on the level of privacy protection they provide. The first category of the Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone [8],[9]. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC [10].However, these algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions that are more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in [8] . For vertically partitioned data, algorithms have been proposed to address the association rule mining [9], k-means clustering [11], and frequent pattern mining problems [12]. The work of [13] uses a secure coprocessor for privacy preserving collaborative data mining and analysis.

The second category of the partial information hiding approach trades privacy with improved performance in the sense that malicious data miners may infer certain properties of the original data from the disguised data. Various solutions in this category allow a data owner to transform its data in different ways to hide the true values of the original data while at the same time still permit useful mining operations over the modified data. This approach can be further divided into three categories: 1) k-anonymity [14], [15], [16], [17], [18], [19], 2) retention replacement (which retains an element with probability p or replaces it with an element selected from a probability distribution function on the domain of the elements) [20], [21], [22], and 3) data perturbation (which introduces uncertainty about individual values before data are published) [1], [2], [3], [4], [5], [6], [7], [23].The data perturbation approach includes two main classes of methods: additive [1], [2], [4], [5], [7] and matrix multiplicative [3], [6] schemes. These methods apply mainly to continuous data. In this paper, we focus solely on the additive perturbation approach where noise is added to data values .Another relevant line of research concerns the problem of privately computing various set related operations. Two party protocols for intersection, intersection size, equijoin, and equijoin size were introduced in [24] for honest-but curious adversarial model. Some of the proposed protocols leak information [25]. Similar protocols for set intersection have been proposed in [26] and [27]. Efficient two party protocols for the private matching problem which are both secure in the malicious and honest-but-curious models were introduced in [28]. Efficient private and threshold set intersection protocols were proposed in [29]. While most of these protocols are equality based, algorithms in [25] compute arbitrary join predicates leveraging the power of a secure coprocessor. Tiny trusted devices were used for secure function evaluation in [30].

.

III. PAPER LAYOUT

The rest of the paper is organized as follows: we go over preliminaries in Section 4. We formulate the problem, and define our privacy goal in Section 5. It highlights the key challenge in achieving our privacy goal, and presents the LI ET AL.: ENABLING ULTILEVEL TRUST IN PRIVACY PRESERVING DATA MINING 1599 intuition that leads to our solution. In Section 6, we formally present our solution, and prove that it achieves our privacy goal.

IV. PRELIMINARIES

A. Jointly Gaussian

In this paper, we focus on perturbing data by additive Gaussian noise [1], [2], [4], [5], [7], i.e., the added noises are jointly Gaussian.

1. Let G1 through GL be L Gaussian random variables. They are said to be jointly Gaussian if and only if each of them is a linear combination of multiple independent Gaussian random variables.

2. Equivalently, G1 through GL are jointly Gaussian if and only if any linear combination of them is also a Gaussian

random variable. A vector formed by jointly Gaussian random variables is called a jointly Gaussian vector.

ISSN: 2278 – 1323

B. *Additive Perturbation*

The single-level trust PPDM problem via data perturbation has been widely studied in the literature. In this setting, a data owner implicitly trusts all recipients of its data uniformly and distributes a single perturbed copy of the data. A widely used and accepted way to perturb data is by additive perturbation [1], [2], [4], [5], [7]. This approach adds to the original data, X, some random noise, Z, to obtain the perturbed copy, Y , as follows:

$Y = X + Z$

C. *Linear Least Squares Error Estimation*

Given a perturbed copy of the data, a malicious data miner may attempt to reconstruct the original data as accurately as possible. Among the family of linear reconstruction methods, where estimates can only be linear functions of the perturbed copy, Linear Least Squares Error (LLSE) estimation has the minimum square errors between the estimated values and the original values [39].estimation is that it simultaneously minimizes all these estimation errors.

V. PROBLEM FORMULATION

In this section, we present the problem settings, describe our threat model, state our privacy goal, and identify the design space. Table 1 lists the key notations used in the paper.

TABLE 1
Key Notations

Notation	Definition
X	original data
Y_i	perturbed copy of X of trust level i
Z_i	noise added to X to generate Y_i
M	number of trust levels
N	number of attributes in X
$\mathbb{Y}$	a vector of all M perturbed copies
$\mathbb{Z}$	a vector of noise Z_1 to Z_M
$\hat{X}(\mathbb{Y})$	LLSE estimate of X given $\mathbb{Y}$
K_X	covariance matrix of X
$K_{\mathbb{Z}}$	covariance matrix of $\mathbb{Z}$

A. *Problem Settings*

In the MLT-PPDM problem, we consider in this paper, a data owner trusts data miners at different levels and generates a series of perturbed copies of its data for different trust levels. This is done by adding varying amount of noise to the data.

Under the multilevel trust setting, data miners at higher trust levels can access less perturbed copies. Such less perturbed copies are not accessible by data miners at lower trust levels. In some scenarios, such as the motivating example we give at the beginning of Section 1, data miners at higher trust levels may also have access to the perturbed copies at more than one trust levels. Data miners at different trust levels may also collude to share the perturbed copies among them. As such, it is common that data miners can have access to more than one perturbed copies. Specifically, we assume that the data owner wants to release M perturbed copies of its data X, which is an N x 1 vector with mean μ_X and covariance Kx as defined in Section 2.2. These M copies can be generated in various fashions. They can be jointly generated all at once .Alternatively, they can be generated at different times upon receiving new requests from data miners, in an on-demand fashion. The latter case gives data owners maximum flexibility. It is true that the data owner may consider to release only the mean and covariance of the original data. We remark that simply releasing the mean and covariance does not provide the same utility as the perturbed data. For many real applications, knowing only the mean and covariance may not be sufficient to apply data mining techniques, such as clustering, principal component analysis, and classification [6]. By using random perturbation to release the data set, the data owner allows the data miner to exploit more statistical information without releasing the exact values of sensitive attributes

VI. SOLUTION TO GENERAL CASES

We now show that the solutions to the general cases of arbitrarily fine trust levels follow naturally from that to the trust levels.

A. *Shaping the Noise*

1. *Independent Noise Revisited*

In Section 4, we show that adding independent noise to generate two differently perturbed copies, although convenient, fails to achieve our privacy goal. The increase in the number of independently generated copies aggravates the situation; the estimation error actually goes to zero as this number increases indefinitely. In turn, the attackers can perfectly reconstruct the original data. We formalize this observation in the following theorem.

2. *Properly Correlated Noise*

We show by the case study that the key to achieving the desired privacy goal is to have noise $Z_i, 1 \le i \le M$ properly correlated. To this end, we further develop the pattern found in the 2 * 2 noise covariance matrix in (13) into a corner-wave property for a multidimensional noise covariance matrix. This property becomes the cornerstone of Theorem 4 which is a generalization of Theorems 1 and 2. Corner-wave Property. Theorem 4 states that for M perturbed copies, the privacy goal in (10) is achieved if the noise

covariance matrix $K_{\mathbb{Z}}$ has the corner-wave pattern as shown in (15). Specifically, we say that an M * M square matrix has the corner-wave property if, for every i from 1 to M, the following entries have the same value as the (I,i) th entry:

- all entries to the right of the (i,i) th entry in row i, and
- all entries below the (i,i) th entry in column i.

The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

B. *Batch Generation*

In the first scenario, the data owner determines the M trust levels priori, and generates M perturbed copies of the data in one batch. In this case, all trust levels are predefined and $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$ are given when generating the noise. We refer to this scenario as the batch generation. We propose two batch algorithms. Algorithm 1 generates noise Z1 to ZM in parallel while Algorithm 2 sequentially.

Algorithm 1. Parallel Generation
1: // Input: X, K_X, and $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$
2: // Output: $\mathbb{Y}$
3: Construct $K_{\mathbb{Z}}$ with K_X and $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$, according to
4: Generate $\mathbb{Z}$ with $K_{\mathbb{Z}}$, according to (16)
5: Generate $\mathbb{Y} = HX + \mathbb{Z}$
6: Output $\mathbb{Y}$

Algorithm 2. Sequential Generation
1: // Input: X, K_X, and $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$
2: // Output: Y_1 to Y_M
3: Construct $Z_1 \sim N(0, \sigma^2_{Z_1} K_X)$
4: Generate $Y_1 = X + Z_1$
5: Output Y_1
6: **for** i from 2 to M **do**
7: Construct noise $\xi \sim N(0, (\sigma^2_{Z_i} - \sigma^2_{Z_{i-1}})K_X)$
8: Generate $Y_i = Y_{i-1} + \xi$
9: Output Y_i
10: **end for**

1. *Algorithm 1: Parallel Generation*

Without loss of generality, we assume $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$ where $1 \le i \le M-1$. Algorithm 1 generates the components of noise $\mathbb{Z}$, i.e., Z1 to ZM, simultaneously based on the following probability distribution function, for any real (N.M)- dimension vector v,

$$f_{\mathbb{Z}}(v) = \frac{1}{\sqrt{(2\pi)^M \det(K_{\mathbb{Z}})}} e^{-\frac{1}{2} v^T K_{\mathbb{Z}}^{-1} v},$$

Where $K_{\mathbb{Z}}$ is given by

Algorithm 1 then constructs $\mathbb{Y}$ as HX -$\mathbb{Z}$ and outputs it. We refer to Algorithm 1 as parallel generation. Algorithm 1 serves as a baseline algorithm for the next two algorithms.

2. *Algorithm 2: Sequential Generation*

The large memory requirement of Algorithm 1 motivates us to seek for a memory efficient solution. Instead of parallel generation, sequentially generating noise Z1 to ZM, each of which a Gaussian vector of N dimension. The validity of the alternative procedure is based on the insight in the following theorem.

3. *Disadvantages*

The main disadvantage of the batch generation approach is that it requires a data owner to foresee all possible trust levels a priori. This obligatory requirement is not flexible and sometimes impossible to meet. One such scenario for the latter arises in our case study. After the data owner already released a perturbed copy Y2, a new request for a less distorted copy Y1 arrives. The sequential generation algorithm cannot handle such requests since the trust level of the new request is lower than the existing one. In today' sever-changing world, it is desirable to have technologies that adapt to the dynamics of the society. In our problem setting, generating new perturbed copies on-demand would be a desirable feature.

C. *On-Demand Generation*

As opposed to the batch generation, new perturbed copies are introduced on demand in this scenario. Since the requests may be arbitrary, the trust levels corresponding to the new copies would be arbitrary as well. The new copies can be either lower or higher than the existing trust levels. We refer this scenario as on-demand generation. Achieving the privacy goal in this scenario will give data owners the maximum flexibility in providing MLT-PPDM services.

VII. CONCLUSION

In this work, we expand the scope of additive perturbation based PPDM to multilevel trust (MLT), by relaxing an implicit assumption of single-level trust in exiting work. MLT-PPDM allows data owners to generate differently perturbed copies of its data for different trust levels. The key challenge lies in preventing the data miners from combining copies at different trust levels to jointly reconstruct the original data more accurate than what is allowed by the data owner. We address this challenge by properly correlating noise across copies at different trust levels. This property offers the data owner maximum flexibility. We believe that multilevel trust privacy preserving data mining can find many applications. Our work takes the initial step to enable MLT-PPDM services. Many interesting and important directions are worth exploring. For example, it is not clear how to expand the scope of other approaches in the area of partial information hiding, such as random rotation-based data perturbation, k anonymity, and retention replacement, to multilevel trust. It is also of great interest to extend our

approach to handle evolving data streams. As with most existing work on perturbation-based PPDM, our work is limited in the sense that it considers only linear attacks. More powerful adversaries may apply nonlinear techniques to

ISSN: 2278 – 1323

derive original data and recover more information. Studying the MLT-PPDM problem under this adversarial model is an interesting future direction.

VIII. REFERENCES

[1] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS '01), pp. 247-255, May 2001.*

[2] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), 2000.*

[3] K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation," *Proc. IEEE Fifth Int'l Conf. Data Mining,2005.*

[4] Z. Huang, W. Du, and B. Chen, "Deriving Private Information From Randomized Data," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2005.*

[5] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, and I. Stanoi, "Hiding in the Crowd: Privacy Preservation on Evolving Streams Through Correlation Tracking," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.*

[6] K. Liu, H. Kargupta, and J. Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," *IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 92-106, Jan. 2006.*

[7] S. Papadimitriou, F. Li, G. Kollios, and P.S. Yu, "Time Series Compressibility and Privacy," *Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), 2007.*

[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," *Proc. Int'l Cryptology Conf. (CRYPTO), 2000.*

[9] J. Vaidya and C.W. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," *Proc. ACM SIGKDD Int'lConf. Knowledge Discovery and Data Mining, 2002.*

[10] O. Goldreichi, "Secure Multi-Party Computation," *Final (incomplete) draft, version 1.4, 2002.*

[11] J. Vaidya and C. Clifton, "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2003.*

[12] A.W.-C. Fu, R.C.-W. Wong, and K. Wang,

[13] "Privacy-Preserving Frequent Pattern Mining across Private Databases," *Proc. IEEE Fifth Int'l Conf. Data Mining, 2005.*

[14] B. Bhattacharjeei, N. Abe, K. Goldman, B. Zadrozny, V.R. Chillakuru, M.del Carpio, and C. Apte, "Using Secure Coprocessors for Privacy Preserving Collaborative Data Mining and Analysis," *Proc. Second Int'l Workshop Data Management on New Hardware (DaMoN '06), 2006.*

[15] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," *Proc. Int'l Conf. Extending Database Technology (EDBT), 2004.*

[16] E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, "Privacy and Ownership Preserving of Outsourced Medical Data," *Proc. 21st Int'l Conf. Data Eng. (ICDE), 2005.*

[17] D. Kifer and J.E. Gehrke, "Injecting Utility Into Anonymized Datasets," *Proc. ACM SIGMOD Int'l Conf. Management of Data,2006.*

[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," *Proc. Int'l Conf. Data Eng., 2006.*

[19] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), vol. 10, pp. 557-570, 2002.*

[20] X. Xiao and Y. Tao, "Personalized Privacy Preservation," *Proc. ACM SIGMOD Int'l Conf. Management of Data, 2006.*

[21] R. Agrawal, R. Srikant, and D. Thomas, "Privacy Preserving OLAP," *Proc. ACM SIGMOD Int'l Conf. Management of Data, 2005.*

[22] W. Du and Z. Zhan, "Using Randomized Response Techniques for Privacy-Preserving Data Mining," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2003.*

[23] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.*

[24] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *Proc. IEEE Third Int'l Conf. Data Mining, 2003.*

[25] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private Databases," *Proc. ACM SIGMOD Int'l Conf.Management of Data, 2003.*

[26] R. Agrawal, D. Asonov, M. Kantarcioglu, and Y. Li, "Sovereign Joins," *Proc. 22nd Int'l Conf. Data Eng. (ICDE '06), 2006. LI ET AL.: ENABLING MULTILEVEL TRUST IN PRIVACY PRESERVING DATA MINING 1611*

[27] C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya, and M. Zhu, "Toolsfor Privacy Preserving Distributed Data Mining," *ACM SIGKDD Explorations, vol. 4, no. 2, pp. 28-34, 2003.*

[28] B.A. Huberman, M. Franklin, and T. Hogg, "Enhancing Privacy and Trust in Electronic Communities," *Proc. First ACM Conf.Electronic*

Commerce, pp. 78-86, Nov. 1999.

[29] M. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," *Advances in Cryptology—EUROCRYPT, vol. 3027, pp. 1-19, 2004.*

[30] L. Kissner and D. Song, "Privacy-Preserving Set Operations," *Proc. Int'l Cryptology Conf. (CRYPTO), 2005.*

[31] A. Iliev and S. Smith, "More Efficient Secure Function Evaluation Using Tiny Trusted Third Parties," *Technical Report TR2005-551, Dept. of Computer Science, Dartmouth Univ., 2005.*

[32] Enabling Multilevel Trust in Privacy Preserving Data Mining *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012*

Mr. Swapnil Kadam received the Bachelor degree (B.E.) in Computer Engineering .Currently, He is pursuing M.E. Computer Engineering at Department of Computer Engg.,BSCOR, Narhe, Pune University. His current research interests include Data mining .

Prof. Navnath B. Pokale obtained M.E. computer. Curently working as a Assistant Professor at epartment of Computer Engg.,BSCOR, Narhe, Pune iversity. He has 14 yrs of teaching experience. His earch interests include Networking, Image ssing and Data Mining.

ISSN: 2278 – 1323