International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

ENHANCING MULTILEVEL TRUST IN PRIVACY PRESERVING DATA MINING WITH ON-DEMAND GENERATION OF TRUST LEVELS

Sujit.S.Ragte
M. Tech Final Year, KBN College of Engg
Gulbarga 585104 Karnataka, India
Email: ssrsujit2010@gmail.com

Dr. Md. Abdul Waheed
Department of Computer Science and Engg
VTU Regional PG Center,
Gulbarga 585106 Karnataka, India

Prof. Shameem Akthar
Department of Computer Science and Engg
KBN College of Engineering,
Gulbarga 585104 Karnataka, India

Abstract: **Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, we relax this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In The Proposed system the more trusted a data miner is the less perturbed copy of the data it can access. Under this system, a malicious data miner may 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 such diversity attacks is the key challenge of providing MLT-PPDM services. We address this challenge by properly correlating perturbation across copies at different trust levels.**

***Index Terms*—Privacy preserving data mining, multilevel trust, random perturbation.**

1. INTRODUCTION:

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining, also called knowledge discovery in databases, in computer sciences, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence such as neural networks and machine learning with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).

Data perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single-level trust on data miners [11]. This approach introduces uncertainty about individual values before data are published or released to third parties for data Mining purposes. Under the single trust level assumption, a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty. This assumption is limited in various applications where a data owner trusts the data miners at different levels.

1.1Fundamental Concepts on (Domain):
Data Mining Overview

Data mining is emerging as one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. To be successful, data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created.

Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Two efforts that have attracted a higher level of congressional interest include the Terrorism Information Awareness (TIA) project (now-discontinued) and the Computer-Assisted Passenger Pre-screening System II (CAPPS II) project (now- cancelled and replaced by Secure Flight).

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. The second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected. A fourth issue is privacy. Questions that may be considered include the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and possible application of the Privacy Act to these initiatives. It is anticipated that congressional oversight of data mining projects will grow as data mining efforts continue to evolve.

1.2 Proposed Enhancement:

On-Demand Generation

As opposed to the batch generation, new perturbed copies are introduced on demand in this second 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 [11].

1.3 Contributions:

- We expand the scope of perturbation-based PPDM to multilevel trust, by relaxing the implicit assumption of single-level trust in existing work. MLTPPDM introduces another dimension of flexibility which allows data owners to generate differently perturbed copies of its data for different trust levels.

- We identify a key challenge in enabling MLT-PPDM services. In MLT-PPDM, data miners may have access to multiple perturbed copies. By combining 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.

- We address this challenge by properly correlating perturbation across copies at different trust levels. We prove that our solution is robust against diversity attacks. We propose several algorithms for different targeting scenarios. We demonstrate the effectiveness of our solution through experiments on real data.

- Our solution allows data owners to generate perturbed copies of their data at arbitrary trust levels. This property offers data owner's maximum flexibility.

- We are also providing the On-Demand generation according to the user's requirement at different trust levels.

2. LITERATURE SURVEY:

1] D. AGRAWAL and C.C. AGGARWAL published paper on "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," in the year MAY (2001).

They developed optimal algorithms and models based on the interesting perturbation approach proposed in R. Agrawal and R. Srikant. Privacy Preserving Data Mining. They proposed a reconstruction algorithm for privacy preserving data mining, which not only converges but does so to the maximum likelihood estimate of the original distribution.

2] Rakesh Agrawal and Ramakrishnan Srikant published a paper on "Information Sharing across Private Databases" in the year 2000.Their Research work included formalizing the notion of minimal information sharing across private databases, and develop protocols for intersection, equijoin, intersection size, and equijoin size.

3] K. CHEN AND L. LIU published a paper on "Privacy Preserving Data Classification with Rotation Perturbation," in the year 2005.

Data perturbation techniques are one of the most popular models for privacy preserving data mining. It is especially convenient for applications where the data owners need to export/publish the privacy-sensitive data.

4] Y. LINDELL AND B. PINKAS published a paper on "Privacy Preserving Data Mining," in the year 2000. In this paper they address the issue of privacy preserving data mining. Specifically, the authors consider as scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information.

3. EXISTING SYSTEM

DATA perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single-level trust on data miners. Under the single level trust assumption a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty.

This approach introduces uncertainty about individual values before data are published or released to third parties for data mining purposes.

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

4. PROPOSED SYSTEM

In this system, 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.

The proposed system allows a data owner to generate distinctly perturbed copies of its data according to different trust levels. It provides a systematic solution to the problem of uncertainty before data is released to third party for data mining purpose. The system allows data owners to generate perturbed copies of their data at arbitrary trust Levels on-demand. The proposed system offer data owners maximum flexibility

We expand the scope of perturbation-based PPDM to multilevel trust, by relaxing the implicit assumption of single-level trust in existing work. MLTPPDM introduces another dimension of flexibility which allows data owners to generate differently perturbed copies of its data for different trust levels. 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. We address this challenge by properly correlating perturbation across copies at different trust levels. We prove that our solution is robust against diversity attacks. We propose several algorithms for different targeting scenarios. We demonstrate the effectiveness of our solution through experiments on real data. Our solution allows data owners to generate perturbed copies of their data at arbitrary trust levels on-demand. This property offers data owner's maximum flexibility.

5. PRELIMINARIES:

5.1 Jointly Gaussian:

In this paper, we focus on perturbing data by additive Gaussian noise the added noises are jointly Gaussian. 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.

5.2 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. This approach adds to the original data, X, some random noise, Z, to obtain the perturbed copy.

5.3 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.

6.IMPORTANT COMPONENTS (MODULES):

6.1 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. [1],[2],[4]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, 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.

6.2 Threat Model :

We assume malicious data miners who always attempt to reconstruct a more accurate estimate of the original data given perturbed copies[2],[3]. We hence use the terms data miners and adversaries interchangeably throughout this paper. In MLT-PPDM, adversaries may have access to a subset of the perturbed copies of the data. The adversaries' goal is to reconstruct the original data as accurately as possible based on all available perturbed copies.

6.3 Privacy Goal and Design Space:

In a MLT-PPDM setting, a data owner releases distinctly perturbed copies of its data to multiple data miners. One key goal of the data owner is to control the amount of information about its data that adversaries may derive.

7. PROPOSED ALGORITHMS:

7.1Batch Generation:

In the first scenario, the data owner determines the M trust levels a 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

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

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 $\sigma^2_{Z_M}$,
4: Generate $\mathbb{Z}$ with $K_{\mathbb{Z}}$, according to
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**

7.2On-demand Generation Algorithm:

Algorithm 3. On Demand Generation

1: // Input: X, K_X, $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$, and values of $\mathbb{Z}'$: v_1
2: // Output: New copies $\mathbb{Z}''$
3: Construct $K_{\mathbb{Z}}$ with K_X and $\sigma^2_{Z_1}$ to $\sigma^2_{Z_M}$, according to
4: Extract $K_{\mathbb{Z}'}$, $K_{\mathbb{Z}''\mathbb{Z}'}$, and $K_{\mathbb{Z}''}$ from $K_{\mathbb{Z}}$
5: Generate $\mathbb{Z}''$ as a Gaussian with mean and variance
6: **for** i from $L+1$ to M **do**
7: Generate $Y_i = X + Z_i$
8: Output Y_i
9: **end for**

8. 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. We have also expanded the work on MLT-PPDM by enhancing the work on On-Demand Generation of Trust levels. We believe that multilevel trust privacy preserving data mining can find many applications. Our work takes the initial step to enable MLT-PPDM services.

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]. Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," Proc. Int'l Cryptology Conf. (CRYPTO), 2000.
[5]. C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Int'l Conf. Extending Database Technology (EDBT), 2004.
[6]. 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.
[7]. C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya, and M. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations, vol. 4, no. 2, pp. 28-34, 2003.
[8]. 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.
[9]. L. Kissner and D. Song, "Privacy-Preserving Set Operations," Proc. Int'l Cryptology Conf. (CRYPTO), 2005.
[10]. X. Xiao, Y. Tao, and M. Chen, "Optimal Random Perturbation at Multiple Privacy Levels," Proc. Int'l Conf. Very Large Data Bases, 2009.
[11]. Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, Published the paper on"Enabling Multilevel Trust in Data Mining "in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, September 2012