IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 11, November 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

A Homomorphic Encryption Approach to Implementing Two-Party Privacy Preserving Data Mining

Dantala O. Oyerinde[1], Bakwa D. Dunka[2] Akese C. Douglas[3]

[1] Department of Computer Science, University Of Jos,
Jos, Plateau State, Nigeria

[2] Department of Computer Science, University of Calabar,
Calabar, Cross Rivers State, Nigeria

[3] Department of Computer Science, University of Jos,
Jos, Plateau State, Nigeria

Abstract

The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithm to leverage this information. This paper is aimed at developing a protocol that will address the issue of privacy in data mining tasks in two party scenarios. We have proposed a framework that uses the homomorphic encryption to add security so that any data mining technique does not lose its valuable data. With the aid of this approach, confidentiality at both parties end is achieved. This model gives valid data mining results for analysis purpose but the actual or true data is not revealed. We discussed implementation evaluation based on metrics proposed by [1], for the framework and algorithm proposed. Tools used include PHP programming language for simulation, MYSQL for database and Visual Paradigm for the modeling analysis. Secondary sources such as academic literature and technical literature from the internet were studied for further classification of processes and techniques.

Keywords: cryptography, confidential databases, Data mining, Data-privacy data, data perturbation, patterns, symmetric, Heuristic.

1. Introduction

Data mining is an emerging field which connects different major areas like databases, artificial intelligence and statistics. Data mining is a powerful tool that can investigate and extract previously unknown patterns from large amounts of data [2]. The process of data mining requires a large amount of data to be collected into a central site. In modern days organizations are extremely dependent on data mining in results to provide better service, achieving greater profit, and better decision-making, for these purposes organizations collect huge amount of data [3]. For example, 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. With the rapid advance of the Internet, networking, hardware and software technology there is remarkable growth in the amount of data that can be collected from different sites or organizations. Huge volumes of Data collected in this manner also include sensitive data about individuals. It is obvious that if a data mining algorithm is run against the union of different databases, the extracted knowledge not only consists of discovered patterns and correlations that are hidden in the data but it also reveals the information which is considered to be private. Privacy is an important issue in many data mining applications that deal with health care, security, financial and other types of sensitive data. 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. Privacy can for instance be threatened when data mining techniques uses the identifiers which themselves are not very sensitive ,but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information The simplest solution to this problem is to completely hide the sensitive data or not to include such sensitive data in the database.

However, this solution is not ideal and accurate because in many applications, like medicine research, DNA research etc. Different organizations or institutions 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 scenario organizations want to share the data but neither of the institute or organizations want to disclose its database or private information about their clients due to privacy concern. In such a situation it is not only necessary to protect 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 special data mining technique which has emerged to protect the privacy of sensitive data and also give valid data mining results [4].

In this paper, we propose a novel framework to preserve the privacy of data owners in data mining. The original data is systematically transformed using randomized data perturbation privacy preserving data mining technique. This tailored data which maintains the characteristics and properties of original data is then submitted as result of customer query through cryptographic techniques.

2 Research Objectives

This paper is aimed at developing a protocol that will address the issue of privacy in data mining tasks in two party scenarios. It specifically considers a 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. We also aim to develop a model/framework that will show how the protocol algorithm can be implemented.

Computers have promised a fountain of wisdom but deliver a deluge of information. This huge amount of data makes it crucial to develop tools (data mining tools) to discover what is called hidden knowledge. The discovered knowledge most at times is sensitive and owners don't accept that the knowledge be exposed to the public and adversaries. This problem motivates research to develop algorithms, techniques and protocols to assure data owners that privacy is protected while satisfying their need to share data for a common good.

3 Literature Review

Research in privacy-preserving data mining has a long history that precedes the popularization of the term data mining." Early work on database privacy examined the problem of databases that answered COUNT and SUM queries provided by users, where COUNT queries returned the number of records which satisfied a predicate, and SUM queries summed over a field for those records which satisfied a predicate. Although these queries in isolation do not reveal individual record values, researchers observed that an adversary could analyze the intersection of several queries in order to tease out sensitive information.

Research in the late 70s and early 80s attempted to solve this problem by either adding noise to the queries, or by restricting the types of queries that are allowed [5]. The definitions of privacy in this early work seem weak by modern standards, since they are only concerned with attacks that reveal a sensitive value with absolute certainty.

Privacy-preserving data-mining research from the late 90s until today has greatly expanded the scope of the database privacy research, by considering different database access scenarios, different types of adversarial abilities, and different privacy goals. There are three main methodologies commonly used in current research, each of which has different and largely incomparable objectives.

The work of [6] (k-anonymity: A model for protecting privacy) is concerned with the re-identification of supposedly anonymous database records by adversaries who link records with identities using quasi-identifiers attributes such as age, sex, and zip code which may be easily found in external databases. Work of this flavor uses the generalization and suppression of quasi-identifiers to make these linkage attacks more difficult.

The work of [7] (Privacy-preserving data mining) is concerned that adversaries may learn the exact values of sensitive attributes, so they add random noise to the data in order to mask true values while still preserving aggregate statistics that are useful to data mining.

The work of [8] (Privacy preserving data mining) examines privacy-preserving data mining from a multi-party computation perspective, in which two parties want to run a data-mining algorithm on a joint database without revealing their own portions of the database to each other. They apply cryptographic techniques from the area of secure multi-party computation to execute data-mining algorithms in a manner that prevents any information other than the algorithm output from being revealed to the parties.

The field of privacy-preserving data mining is a vast landscape encompassing a wide variety of problems, each of which has different trust models, goals, and scenarios. In this section, I examine work on several popular privacy-preserving data-mining problems and techniques. The amount of data that need to be processed to extract some useful information is increasing. So the methods used for extracting information from huge amount of data must be optimum. As described, the various data mining algorithms can be classified into two broad categories [9].

1. Heuristic-based approaches
   - Additive noise
   - Multiplicative noise
   - K-Anonymization
   - Statistical disclosure control based approaches

2. Cryptography -based approaches

3.1 Additive-Noise-based Perturbation Techniques

Random noise is added to the actual data in additive-noise-based perturbation technique. The privacy is measured by evaluating how closely the original values of a modified attribute can be determined. In particular, if the perturbed value of an attribute can be estimated, with a confidence c, to belong to an interval [a, b], then the privacy is estimated by (b−a) with confidence c. However, this metric does not work well because it does not take into account the distribution of the original data along with the perturbed data.

3.2 Multiplicative-Noise-based Perturbation Techniques

Additive random noise can be filtered out using certain signal processing techniques with very high accuracy. This problem can be avoided by using random projection-based multiplicative perturbation techniques. Instead of adding some random values to the actual data, random matrices are used to project the set of original data points to a randomly chosen lower-dimensional space. However, the transformed data still preserves much statistical aggregate regarding the original dataset so that certain data mining tasks can be performed on the transformed data in a distributed environment (data are either vertically partitioned or horizontally partitioned) with small errors. High degree of privacy of original data is ensured in this approach. Even if the random matrix is disclosed, it only approximate value of original data can be estimated. It is impossible to get back the original data. The variance of the approximated data is used as privacy measure.

3.3 k- Anonymization Techniques

K-anonymization technique for privacy preservation was introduced by Samarati and Sweeney. A database is k-anonymous with respect to quasi-identifier attributes if there exist at least k transactions in the database having the same values according to the quasi-identifier attributes. In practice, in order to protect sensitive dataset T, before releasing T to the public, T is converted into a new dataset T* that guarantees the k-anonymity property for a sensible attribute. This is done by generalizations and suppression on quasi-identifier attributes. Therefore, the degree of uncertainty of the sensitive attribute is at least 1/k.

3.4 Statistical-Disclosure-Control-based Techniques

To anonymize the data to be released (such as person, household and business), which can be used to identify an individual, additional information publicly available need to be considered. Among these methods specifically designed for continuous data, the following masking techniques are described: additive noise, data distortion by probability distribution, resampling, rank swapping, etc. The privacy level of such method is assessed by using the disclosure risk, that is, the risk that a piece of information be linked to a specific individual.

3.5 Cryptography-based Techniques

The cryptography-based technique usually guarantees very high level of data privacy. Generally solution is based on the assumption that each party first encrypts its own item sets using commutative encryption, then the already encrypted item sets of every other party. The two communicating party must share a common key which is used for encryption and decryption. Sometimes two key is used known as public key and private key. Public key is known to everybody that wants to communicate with you and private key is used for decryption in a secure communication. Though cryptography-based techniques can well protect data privacy, they may not be considered good with respect to other metrics like efficiency.

The utility of the data must be preserved to a certain extent at the end of the privacy preserving process, because in order for sensitive information to be hidden, the database is essentially modified through the changing of information (through generalization and suppression) or through the blocking of data values. Sampling is a privacy preserving technique which does not modify the information stored in the database, but still, the utility of the data falls, since the information is not complete in this case. As we go on changing the data for preserving privacy, the less the database reflects the domain of interest. So, one of the evaluation parameter for the measuring data utility should be the amount of information that is lost after the application of privacy preserving process. Of course, the measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed. As defined in [10], information loss in the context of association rule mining will be measured either in terms of the number of rules that were both remaining and lost in the database after sanitization, or even in terms of the reduction/increase in the support and confidence of all the rules. For the case of classification, we can use metrics similar to those used for association rules. Finally, for clustering, the variance of the distances among the clustered items in the original database and the sanitized database can be the basis for evaluating information loss in this case.

The Paper: A General Survey of Privacy Preserving Data Mining Models and Algorithms [11], identified the following important drivers as the main reasons why privacy preserving data mining has tremendous potentials. The important drivers are:

i. Web demographics

ii. Inter-Enterprise data mining

iii. Security applications

From the review of the literature, it was observed that the field of privacy-preserving data mining has two approaches to the problem of executing machine-learning algorithms on private data. One approach sanitizes the data through suppression and generalization of identifying attributes and/or addition of noise to individual data entries. The sanitized version is then published so that interested parties can run any data-mining algorithm on it.

The other approach is to use cryptographically secure multi-party computation techniques to construct protocols that compute the same answer as would have been obtained in the non-private case. This approach has typically been applied when the relationship between the parties is symmetric: for example, the database is partitioned between them and the result of the protocol execution is that both parties learn the same output based on the joint database. By contrast, in the sanitization approach, the parties executing the data-mining algorithms do not have any data of their own, while the database owner obtains no output at all.

Privacy-preserving data mining remains a difficult problem. Real-world solutions will need to rely on a combination of legal, regulatory, and technological components. It is unlikely that we will ever reach a point where technological solutions alone can completely guarantee the privacy of individuals while allowing for meaningful exploratory data mining. Nevertheless, algorithms such as those developed in the various thesis highlighted above remain useful because they precisely identify what security guarantees are possible under different scenarios. This provides a framework that allows legal regulations to be more precisely followed when they are deployed in actual software to be used in real-world scenarios.

4 Research Methodology

In our approach, we implement privacy preservation in data mining by using the homomorphic encryption to add security so that any data mining technique does not lose its valuable data. We used the asymmetric encryption with RSA encryption where we assumed that decryption occurs only at the data owner's (party 1 and 2) domain.

In this framework the total process is divided into three components the two database owners, and a trusted third party (robot). The role of the third party is purely passive; it keeps the record of no. of data providers, runs data mining algorithms on union of databases and transfers the results/rules/patterns back to the data owners and other data providers.

4.1 Proposed Algorithm

Step 1: A party (party one) initiates a mining process by reaching out to a second party (party two).
Step 2: Upon second party's agreement, both parties jointly contract a trusted third party.
Step 3: The trusted third party sends its public key to both parties.
Step 4: Both parties encrypt their input data with their public keys and send to the trusted third party.
Step 5: The trusted third party decrypts the data using its private key and runs the desired mining algorithm on the data.
Step 6: The trusted third party encrypts the output/results of the mining algorithm with parties' public keys and send back to them.
Step 7: The parties decrypts the output using their private keys and obtain the mining outputs in a consolidated form.

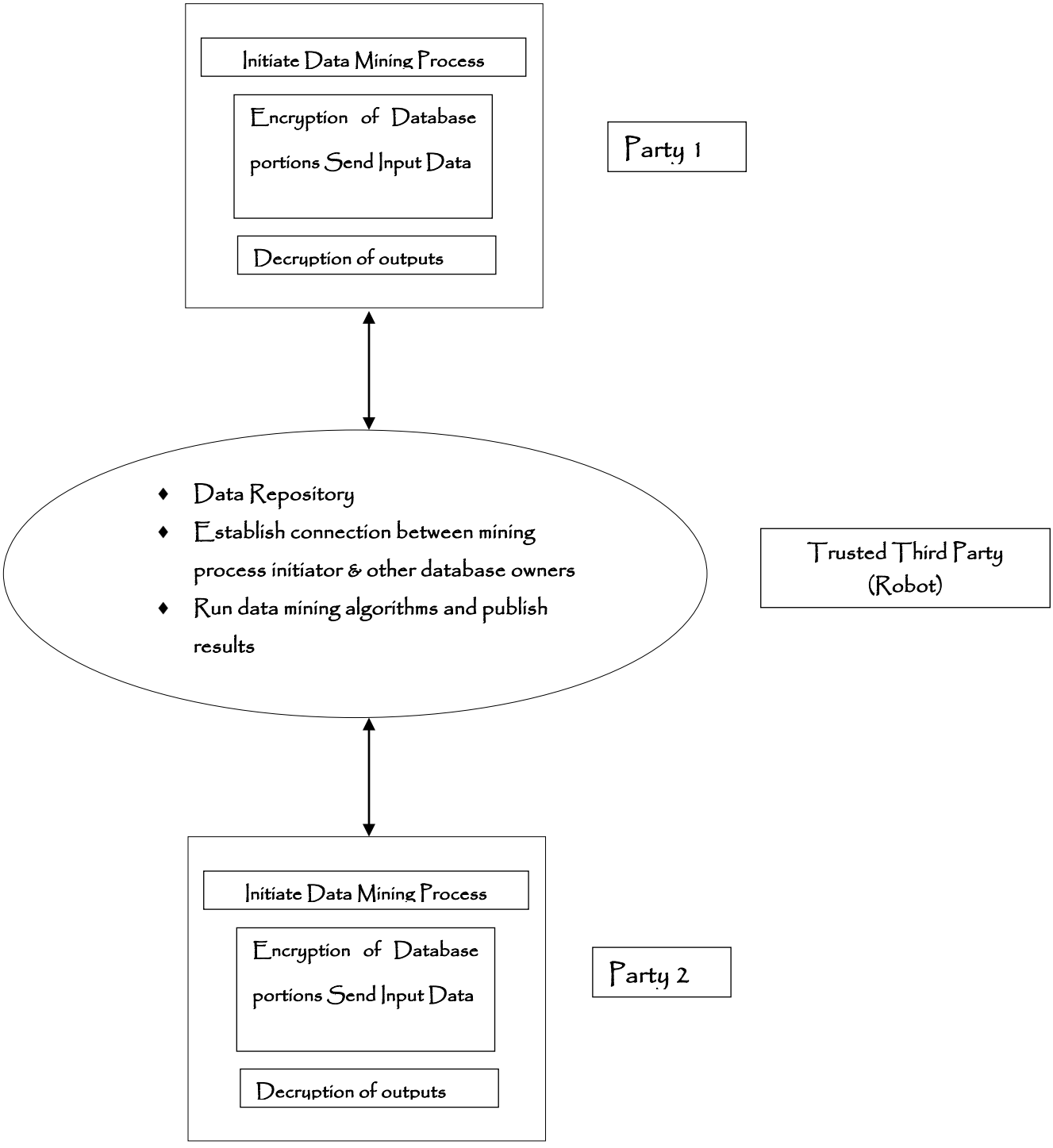

Fig. 1: Framework for privacy preservation Data mining

5 Analysis and Design

A major problem of existing systems is that of not been able to preserve the privacy of individuals as obtained from database owners and in situations where little privacy exist, data is seriously distorted.

In an attempt to solve the above problem, we have proposed a framework that will preserve privacy in data mining scenarios where two or more parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information, the proposed system will concentrate on Creating encryption and decryption algorithms (framework) that ensure the privacy of private inputs.

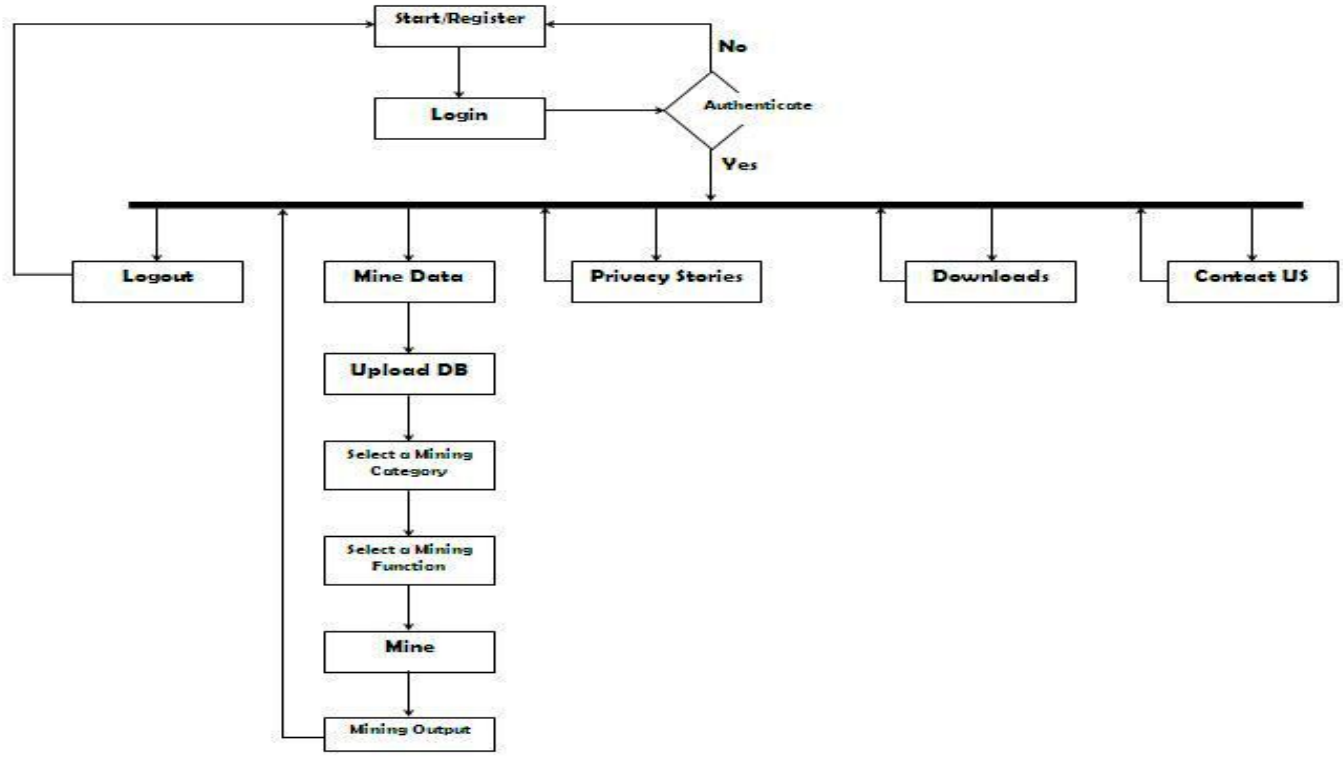

Fig. 2: Context Model of Proposed System

6 Discussions

Given the number of different privacy preserving data mining (PPDM) techniques that have been developed in recent years, there is an emerging need of moving toward standardization in this new research area [12]. One step toward this essential process is to provide a quantification approach for PPDM algorithms to make it possible to evaluate and compare such algorithms. However, due to the variety of characteristics of PPDM algorithms, it is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria.

Rather, an algorithm may perform better than another one on specific criteria like privacy level, data quality. Therefore, it is important to provide users with a comprehensive set of privacy preserving related metrics

which will enable them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are interested in optimizing [13].

[14] discussed in length some quantifying metrics which can be used for PPDM algorithm evaluation. We shall use these metrics for the evaluation discussion of our proposed PPDM algorithm. These metrics are:

- *Privacy level* offered by a privacy preserving technique, which indicates how closely the sensitive information, that has been hidden, can still be estimated.
- *Hiding failure*, that is, the portion of sensitive information that is not hidden by the application of a privacy preservation technique.
- *Data quality* after the application of a privacy preserving technique, considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied;
- *Complexity*, that is, the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm.

Privacy levels are generally classified into data privacy and result privacy**.** The cryptography-based technique guarantees very high level of data privacy. The solution is based on the assumption that each party first encrypts its own item-sets using commutative encryption, then the already encrypted item-sets of every other party. Later on, an initiating party transmits its frequency count, plus a random value, to its neighbor, which adds its frequency count and passes it on to other parties. Finally, a secure comparison takes place between the final and initiating parties to determine if the final result is greater than the threshold plus the random value.

Our framework considers three types of data: public data (P), accessible to everyone including the adversary; private/sensitive data (S), must be protected and unknown to the adversary; unknown data (U), not known to the adversary, but the release of this data might cause privacy violation. The framework assumes that S depends only on P and U, and the adversary has at most *t* data samples of the form (*pi*, *si*). In line with [15], the approach to determine whether an inference channel exists is comprised of two steps. First, a classifier $C1$ is built on the t data samples. To evaluate the impact of C, another classifier $C2$ is built based on the same t data samples plus the classifier C. If the accuracy of $C2$ is significantly better than $C1$, we can say that C provides an inference channel for S. Classifier accuracy is measured based on Bayesian classification error. Suppose we have a dataset

$\{x1, \ldots, xn\}$, and we want to classify xi into m classes labeled as $\{1, \ldots, m\}$. Given a classifier C: $C : xi \rightarrow C(xi) \in \{1, \ldots, m\}$, $i = 1, \ldots, n$

The classifier accuracy for C is defined as:

$$\sum_{j=1}^{m} Pr(C(xi) = j | z = j) Pr(z = j)$$

Where z is the actual class label of xi

Since cryptography-based PPDM techniques usually produce the same results as those mined from the original dataset, analyzing privacy implications from the mining results is particular important to this class of techniques.

For quantifying hiding failure, the percentage of sensitive information that is still discovered, after the data has been sanitized, gives an estimate of the *hiding failure* parameter. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure. Thus, they hide all the patterns considered sensitive.

However, it is well known that the more sensitive information we hide, the more non-sensitive information we miss. Thus, some PPDM algorithms have been recently developed which allow one to choose the amount of sensitive data that should be hidden in order to find a balance between privacy and knowledge discovery. For example, [16], define the *hiding failure* (HF) as the percentage of restrictive patterns that are discovered from the sanitized database. It is measured as follows:

$$HF = \frac{\#R_P(D')}{\#R_P(D)}$$

Where $\#RP(D)$ and $\#RP(D')$ denotes the number of restrictive patterns discovered from the original data base D and the sanitized database D' respectively. Ideally, HF should be 0. In their framework, they give a specification of a ϕ, representing the percentage of sensitive transactions that are not sanitized, which allows one to find a balance between the hiding failure and the number of misses.

Note that ϕ does not control the *hiding failure* directly, but indirectly by controlling the proportion of sensitive transactions to be sanitized for each restrictive pattern.

For quantifying data quality, the main feature of most PPDM algorithms is that they usually modify the database through insertion of false information or through the blocking of data values in order to hide sensitive information. Such perturbation techniques cause the decrease of the data quality. It is obvious that the more the changes are made to the database, the less the database reflects the domain of interest. Therefore, data quality metrics are very important in the evaluation of PPDM techniques. Since the data is often sold for making profit, or shared with others in the hope of leading to innovation, data quality should have an acceptable level according also to the intended data usage. If data quality is too degraded, the released database is useless for the purpose of knowledge extraction [17].

In existing works, several data quality metrics have been proposed that are either generic or data-use-specific. However, currently, there is no metric that is widely accepted by the research community [18]. In evaluating the data quality after the privacy preserving process, there is need to assess both the *quality of the data* resulting from the PPDM process and the *quality of the data mining results*. The quality of the data themselves can be considered as a general measure evaluating the state of the individual items contained in the database after the enforcement of a privacy preserving technique. The quality of the data mining results evaluates the alteration in the information that is extracted from the database after the privacy preservation process, on the basis of the intended data use.

7 Conclusion

Research area of the data mining Privacy preserving is an ongoing research area and there are lots of issues that need to be addressed. In our approach, we have implemented privacy preservation in data mining by using the homomorphic encryption to add security so that any data mining technique does not lose its valuable data. We used the asymmetric encryption with RSA encryption where we assumed that decryption occurs only at the data owner's (party 1 and 2) domain.

The concept of data mining has been around for long time but it took the innovative computing technology and software of the last decade for it to develop into the effective tool it is nowadays. Data mining is a powerful tool but like all powerful things is subject to abuse, misuse and ethical considerations. To ensure the integrity of its use, and therefore the confidence of the users, research must adequately regulate itself concerning privacy issues. Failure to do so will increase the hesitation of individuals as well as organizations from releasing or exchanging data which will affect the performance of these organizations and limit their ability to take steps for the future, not to mention that the release of sensitive data will invite intervention of the authorities, which will create its own set of problems.

8 Further Research

Certain elements in this paper leave scope for further development. With almost any research in current Information Systems/Science/Technology, a list of future enhancements could be endless. In this case, we will only highlight the general areas where extra work would benefit the project.
An immediate extension of this work is to extend to the case of many (rather than two) parties. A problem which arises is that our protocol does not extend to the case of many parties.

A very open problem is to design an efficient protocol which is secure even when one of the parties is malicious and does not necessarily act according to the protocol. Other future work includes considering other important data mining algorithms such as Neural Networks and Association rule Mining. The use of elliptical cryptography can also be considered in the future. Alternatively, a formal framework can be developed that upon testing of a PPDM algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of PPDM algorithms.

9 References

[1],[11],[14],[15],[17],[18] Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining* (pp. 11-52). Springer US.

[2] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques" In Proc. of 3rd IEEE Int. Conf. on Data Mining, Washington, DC, USA,, pages99–106, 2003.

[3] Agrawal R.,Srikant R., "Privacy Preserving Data Mining.," In the Proceedings of the ACM SIGMOD Conference. 2000. K.Muralidhar.,R.Sarathi, "A General additive data perturbation method for data base security" journal of Management Science. 45(10):1399-1415,2002

[4] Clifton, C, Kantarcioglu, M, Vaidya,J Lin, X, Zhu, M Y. ,

"Tools for privacy preserving distributed data mining" , SIGKDD Explor. Newsl., 28-34, 2002

[5] K. Muralidhar, R. Sarathy, and R. A. Parsa, "A general additive perturbation method for database security" Management Science, vol. 45, no. 10, pp. 1399–1415, 1999.

[6]Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 571-588.

[7] 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

[8] Lindell Y., Pinkas, B."Privacy preserving Data Mining"CRYPTO 2000.

[9],[13] Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, *11*(2), 121-154.

[10] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, *33*(1), 50-57.

[12],[16] Oliveira, S. R., & Zaïane, O. R. (2004). Toward standardization in privacy-preserving data mining. In *ACM SIGKDD 3rd Workshop on Data Mining Standards* (Vol.7).

First Author
Oyerinde Dantala Oyeyinka received a B.Sc. (Honours) degree in Computer Science (Information Systems) from Babcock University, Ilishan Remo, Ogun State, Nigeria in 2005. He received his M.Sc. degree in Computer Science (Management Information Systems) from the same Babcock University in 2014. He is currently a lecturer in the Department of Computer Science, University of Jos, Jos Nigeria. His current research interests include Data Analytics, Learning Analytics, Big Data, Virtualization and Distributed Computing, Knowledge Management and eLearning. He is a member of the Nigerian Computer Society (MNCS), Association for Information Systems (AIS) and the Internet Society (ISOC).

Second Author
Bakwa, Dunka Dirting received a B.Sc. (Honors) degree in Computer science from the University of Jos, Plateau state, Nigeria in 2005, He recently concluded an M.Sc. degree in Computer science from the University of Calabar, Cross Rivers State, Nigeria. October, 2016. He is currently a System Programmer with the Corporate Information Systems, ICT Directorate, University of Jos, where he worked on various Application development and system maintenance project for both students and staff. His main areas of interest in the field of academics and Computer Science research include Information Systems and Knowledge Management, Embedded and Real-Time Systems, Human Computer Interaction, Software Engineering and Programming Languages, Data Sciences and Analytics.

Third Author
Akese Douglas is currently pursuing a B.Sc. degree in Computer Science with the University of Jos, Jos Nigeria.