Incorporating Functional Encryption(FE) in Privacy Preserving Data Mining(PPDM)

Submitted By

Yaman Patel

13MCEI14

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Incorporating Functional Encryption(FE) in Privacy Preserving Data Mining(PPDM)

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering
(Information and Network Security)

Submitted By

Yaman Patel

(13MCEI14)

Guided By

Dr. Sanjay Garg, Dr. Sharada valiveti

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Certificate

This is to certify that the major project entitled **"Incorporating Functional Encryption(FE) in Privacy Preserving Data Mining(PPDM)"** submitted by **Yaman Patel (Roll No: 13MCEI14)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering(Information and Network Security) of, Institute of Technology, Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr. Sanjay Garg
Guide & Professor & Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sharada valiveti
Guide & Associate Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Yaman Patel**, Roll. No. **13MCEI14**, give undertaking that the Major Project entitled "**Incorporating Functional Encryption(FE) in Privacy Preserving Data Mining(PPDM)**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student

Date:

Place:

Endorsed by

Dr. Sanjay Garg, Dr. Sharada Valiveti

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr. Sanjay Garg, Dr. Sharada Valiveti**, Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Yaman Patel
13MCEI14

Abstract

Privacy preserving data mining has emerged due to large usage of data in organizations for extracting knowledge from data. Big data uses centralized as well as distributed data and mines knowledge. Privacy preservation of data has become critical asset due to malicious users and society issues. This paper expresses issues in privacy preserving data mining which includes both cryptographic and non-cryptographic approaches. Due to security concern, cryptographic approaches like Homomorphic encryption, Shamir's secret sharing schemes and oblivious transfers are more focused. Usage of these approaches increases communication and computation cost of data mining operations obviously. This paper has incorporated new approach in privacy preservation, Functional Encryption (FE). FE uses personalized randomness, bi linear groups for cryptographic key mapping, permutations etc. which makes it more complex, but more efficient. Two algorithms are proposed with Trusted Third Party and Collaborative processing model incorporating FE schemes. FE provides higher level of security and data privacy. FE only allows to learn the output of function without revealing anything else. Final model exhibits feasible computation cost. Communication cost in (Semi Trusted Authority) STA model is $O(n^3)$, while in (Semi Trusted Third Party) STTP model, it is $O(n^2 logn)$.

Abbreviations

PPDM	Privacy Preserving Data Mining.
PPDDM	Privacy Preserving Distributed Data Mining.
FE	Functional Encryption.
IBFE	Identity Based Functional Encryption.
ABFE	Attribute Based Functional Encryption.
CCA	Choosen Ciphertext Attack.

–

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Privacy preserving data mining has emerged due to large usage of data in organizations for extracting knowledge from data[1]. Big data uses centralized as well as distributed data and mines knowledge. Privacy of data has become critical asset due to malicious users and society issues. It is very crucial nowadays to maintain balance between ensuring privacy and extracting knowledge. This areas is burning domain for researchers till now because no any research has been done that out perform all the technique in privacy preserving data mining. Privacy preservation is classified into many categories like data modification, data distribution, data hiding and data encryption. For performance measuring evaluation criteria like information loss, computational overhead, data utility etc are considered[3]. Data modification techniques mainly focusing on adding errors to data or results output which degrades the accuracy of data mining algorithm. In case of critical analysis of data, we have adopted crypto graphical approaches in privacy preserving data mining which has no loss of information but overhead of computation and communication. PPDM includes homomorphic encryption, Shamirs secret sharing scheme, oblivious transfer and many other cryptography techniques. Challenges in this area includes higher computational and communication cost.

In todays information age, data collection is ubiquitous, and every transaction is recorded somewhere. The resulting data sets can consist of terabytes or even petabytes of data, so efficiency and scalability is the primary consideration of most data mining algorithms

Naturally, ever-increasing data collection, along with the goal of data mining is to extract knowledge from data leads to privacy concerns. User friendliness of data mining results leads us to protect against leakage of individuals private information. for example with the help of join operation on databases which are private and publicly available, private information about any citizen can be leaked. Data can be homogeneous or it can be heterogeneous in nature. In distributed environment, where risk is higher in processing data for mining as it has more number of issues like secure line for communication, honest parties involved, third party behavior etc. Higher level of security is needed to overcome such issues. Due to higher level of security and privacy provided by cryptography based approaches, we apply these to provide privacy as well as security. In privacy preserving distributed data mining, two types of communication models are used. Trusted third party and collaborative processing. In case of trusted third party, all the computation and key distribution party in handled by central authority while in later model, parties themselves takes care of aggregation of results. later approach exhibits higher cost due to higher computation. Advanced development of cryptography like functional encryption has proven significantly helpful in data privacy. Though it is based on many assumptions of cryptography, it is highly secure and efficient scheme till now in cryptography. This work have analyzed the concept of Functional Encryption[46] like Attribute based Encryption, Identity based Encryption and Inner Product based Encryption. Two model for incorporating functional encryption in privacy preserving distributed data mining have been proposed. Functional Encryption allows receiver to learn the output of the function defined instead of data which is encrypted. we are trying to use the concepts in our privacy preserving model.

1.1 Motivation

In privacy preserving distributed data mining, knowledge discovery is done by processing the data resided in remote sites. Every database struggles against privacy and security scenarios. Privacy preserving distribute data mining provides both privacy and security to knowledge discovery process.

Privacy preserving data mining consists of two approaches, the modification based and cryptography based approaches. Later provides higher level of privacy as well as security in terms of distributed scenario. Former approach is useful in scenarios where data is at one place and only modification of results or data ensure the privacy factor. We will focus on later approach as it provides both security and privacy but doesnt have any standardize framework and have higher computational and communicational overheads.

Cryptography based approaches has many flavors, we are focusing on following techniques which suitable to our case. Homomorphic encryption based, oblivious transfer, elliptic curve cryptography based, secret sharing based technique are to ensure the secrecy and privacy of the database. Oblivious transfer has higher communicational and computational overheads and they are not suitable for larger databases. So our focus will be on remaining techniques to be applied in our research. We use this technique along with distributed data mining algorithms for best results of our research. Homomrphic encryption has two approaches, symmetric encryption and asymmetric encryption. Later is very useful in distributed scenarios as it has separate mechanisms in case of third party exists, former approach has lower overheads but sometimes is fails to provide security factor. Secret sharing has approaches like shamirs secret sharing scheme and verifiable secret sharing scheme. Both has their advantages and disadvantages need to be worked upon.

We consider two scenarios for processing the data, third party presence, collaborative processing. Both needs to be verified for their adversarial behavior. Advanced study in cryptography field shows positive and improved results in Functional encryption scheme for data privacy. We will also use concept of functional encryption in our framework. Our main focus in providing security to processing by reducing the overheads and application to real world scenario of our research.

1.2 Objectives

Recent research in privacy preserving data mining is burning topic due to remotely located database, privacy laws and friendliness of data mining results. Privacy preserving data mining has many approaches needs to be explore to achieve qualitative and innovative results. At this age of big data, privacy concern has grown more for individuals as well as for organizations. When it comes to privacy providing, security that comes into our mind. Security increases cost which leads to degradation of overall scenario. Our objective is to achieve qualitative research results in distributed environment by applying data mining algorithm with provision of security. We try to reduce the cost of security. We have adopted approaches like homomorphic encryption, secret sharing schemes, oblivious transfer. In earlier stage, we have analyzed all the techniques of security provision with data mining algorithm. We have found out the flaws and limitation. With that information, we have produced framework of privacy preserving distributed data mining. At the end, we will implement proposed framework and produce the results.

Our main objective for research are as follows:

- Analysis of encryption techniques incorporation with data mining algorithms in distribute environment.
- Framework for working in privacy preservation in distributed environment
- Implementation of privacy preserving distributed data mining.
- Efficient and quality results production and real world application solutions with results.

Chapter 2

Literature Survey

In field of cryptography, researchers have given many fruitful results in last decades. But due to increasing demand of privacy for mining results, Researchers till now trying to give there best solutions to the issues. We have seen many firm solutions to this field and mention below in our survey.

Table 2.1: Heuristics based

AUTHOR	YEAR	APPROACH	FINDINGS
Samarati P.	2001	First annonymization	Usage of generalization and suppression. Provides identity disclosure.
Olivieria Zalane	2002	Usage of sanitizing data	Only applicable to association rule mining.
Zhong S. et. al.	2005	Heuristic and cryptography based solution	Provides end to end security.

Table 2.2: Reconstruction based

AUTHOR	YEAR	APPROACH	FINDINGS
Srikant and Agrawal	2001	First technique proposed	Usage with classification approach
Dutta H. et. Al.	2003	Data distortion using noise	Smaller noise gives good result
Kamrakar and Bhattacharya	2009	Randomization and perturbation to modify data	Works good with centralized data. Biased towards privacy at the cost of data utility
Xiaolin and Honglin	2010	Amplifying matrix condition used	Better trade between usability and privacy

Table 2.3: cryptography based technique

AUTHOR	YEAR	APPROACH	FINDINGS
Lindell and Pinkas[35]	2000	Secure multiparty computation	On ID3 algorithm. Higher computational cost
Vaidya and Clifton[18]	2003	Vertically partitioned data, clustering algorithm	Tradeoff between comp. cost and privacy
Bing Yang, Hiroshi, nakagawa, jun sukama[26]	2010	Secure summation, homomorphic encryption	Compared to vaidyas protocol communication cost is low
Sankita and d c jinwala[16]	2013	Shamirs secret sharing scheme, collaborative model	Reduced cost of computation.

Table 2.4: Functional Encryption Research

AUTHOR	YEAR	APPROACH	FINDINGS
Naveed, Shashank, Manoj [39]	2014	controlled functional encryption, Inner product construction	Framework for data privacy based on cryptography assumptions
Shashank, Shweta[40]	2013	IND and SIM based FE definition, data privacy	Selectively secure FE against FE
Allison, Tatsuaki[45]	2011	KP & CP based ABE, bi linear map, Inner Product based	Fully secure in dual encryption method
Barbosha, Farshim[42]	2012	Verifiable computing, strong secure DHE	Works in non-adaptive, bounded environment, it cant provide function privacy
Sergey, Vinod, Hoeteck[43]	2012	FE with Bounded collusion via Multiparty Computation	Q bounded, non adaptive, selective secure

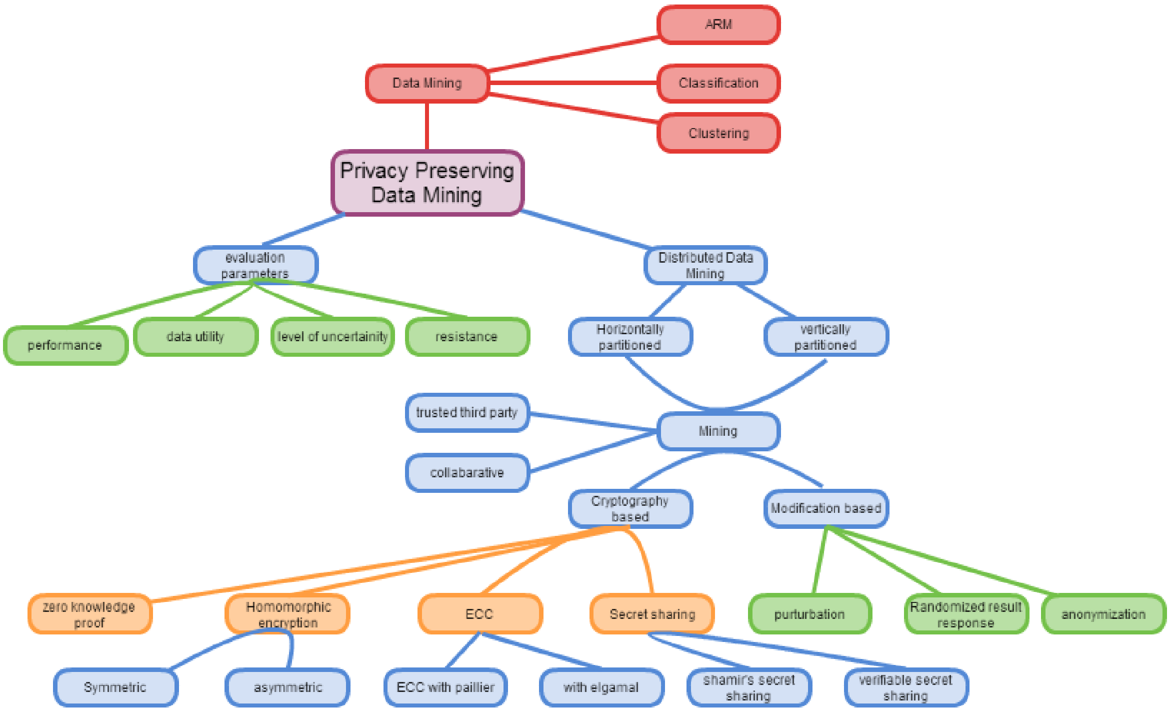

Figure 2.1: Literature Survey

Chapter 3

Privacy Preserving Data Mining(PPDM)

Information mining and learning revelation in databases are two new research territories that explore the programmed extraction of awhile ago obscure examples from a lot of information[10]. In this new age, due friendliness data mining results, researchers have considered the aspect of privacy in mining algorithms. It is decently archived that this new without breaking points blast of new data through the Internet and other media, has arrived at to a point where dangers against the protection are exceptionally basic once a day and they merit genuine considering. Protection protecting information mining[4] is a novel exploration heading in information mining and measurable databases, where security and privacy both are needed to be analyzed. we consider following two points for PPDM. First, sensitive data like identifiers, names, addresses and the like, should be removed from the main database because miner of the data not to be able to compromise another persons privacy[6]. Second, sensitive information which can be mined from a database by using data mining algorithms, should also be trimmed, because such a knowledge can equally well compromise data privacy which is our purpose. The principle target in privacy preserving data mining[4] is to create calculations for altering the first information somehow, so that the protection of information and concentrated learning stay private significantly after the mining methodology. The issue that emerges when private information can be gotten from discharged information by unapproved clients is likewise usually called the database induction issue.

3.1 Approaches in Privacy Preserving Data Mining

Following figure illustrates approaches in privacy preserving data mining.

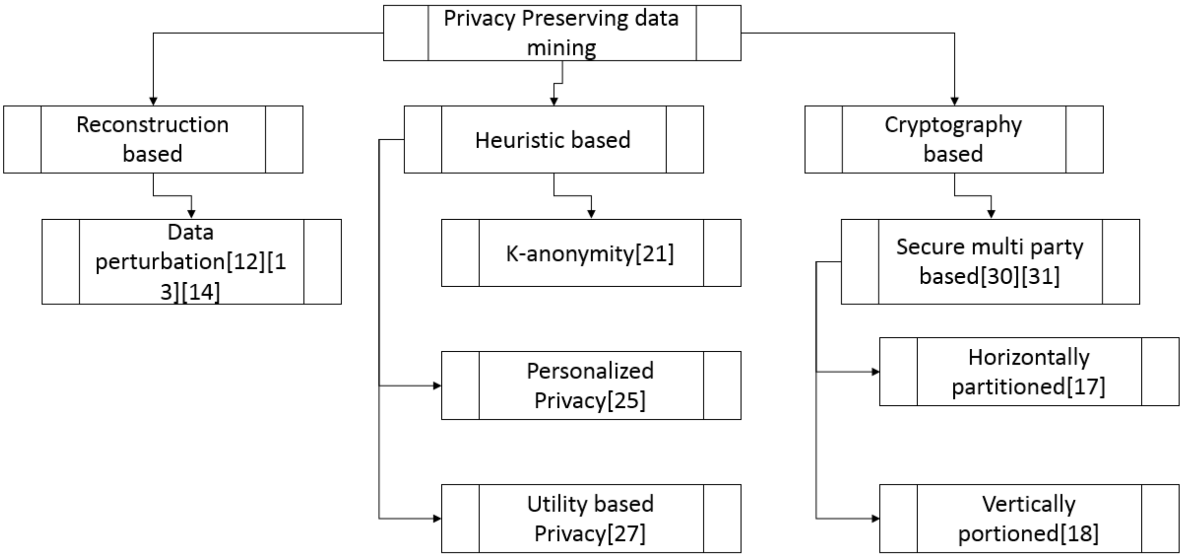
Privacy Preserving data mining
Reconstruction based
Heuristic based
Cryptography based
Data perturbation[12][13][14]
K-anonymity[21]
Secure multi party based[30][31]
Personalized Privacy[25]
Horizontally partitioned[17]
Utility based Privacy[27]
Vertically portioned[18]

Figure 3.1: Privacy preservation approaches

3.1.1 Anonymization technique

In the most basic form of PPDM, following is the table format given by data owner

D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes)

Explicit Identifier is a set of characteristics, for example, name and standardized savings number, containing data that expressly distinguishes record managers[25]. Quasi Identifier is a situated of qualities that could conceivably distinguish record holders. Sensitive Attributes comprises of delicate individual particular data, for example, sickness, pay, and handicap status and Non-Sensitive Attributes contains all properties that don't fall into the past three classes. The four sets of properties are disjoint restricted. Anonymization[2] refers to the PPDM approach that seek to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed[25].

3.1.2 Data perturbation

This techniques is classified into two categories 1) probability distribution sampling 2) fixed data perturbation. The probability distribution type considers the database to be an example from a given data that has a given probability distribution. In given situation security controlling technique replaces the first information by an alternate example from the same ditribution or by the appropriation itself. In the fixed data perturbation type, the estimations of the characteristics in the database, which are to be utilized for processing insights, are perturbed once and for all. The altered information perturbation strategies have been created solely for either numerical information or absolute information.[13]. Following are Data perturbation methods:

- *Noise additive perturbation*, generally adds additive noises to attribute values. The data owner may not need all the column to be private so column based perturbation is applied. This technique is simplest form of perturbation which includes additive perturbation and additive perturbation.

- *Condensation-based* perturbation aims at preserving covariance matrix for multiple columns.

- *Random projection based* perturbation, refers to the technique of projecting a set of data points from the original multidimensional space to another randomly chosen space. Let Pkd be a random projection matrix, where Ps rows are orthonormal [15].

3.1.3 Cryptography Based Technique

Investigate in protection saving information mining began after 2000, yet the cryptographic foundation goes over to Yao's definition and answer for the ”tycoons issue” in 1982[8].In Yao's moguls issue two tycoons need to figure out who is wealthier, yet without uncovering their wealth to the one another. Actually, the capacity to look at numerical information is essential in most information mining assignments. Yao's work launched research in secure multi-party processing which is the investigation of the class of capacities where two or more players can safely process on their joint inputs. This is carried out in a manner that only the last consequence of the reckoning will be uncovered to the gatherings. Specifically, no gathering will know the inputs of alternate gatherings. Mainly privacy preserving distributed data mining uses cryptography techniques to gather local models from local sites and aggregate into central model. This techniques helps in accuracy of results but degrades the performance of data mining algorithm as the computational overhead in higher due to complexity of cryptography algorithm. Shamirs secret sharing[7], homomorphy encryption[11], secure multi party protocols[12], public key cryptosystems etc. algorithms are there in privacy preserving data mining.

Public key encryption plans are focused around higher computational expense and they oblige methods, for example, modular exponentiation of large numbers (in the order of 1k bits). on the contrary, we trust it is extremely productive to figure secret shares when utilizing e.g. Shamir's secret sharing or the simple additive secret sharing. Public key encryption system create cipher texts of at least 1024 bits. If we try to use the homomorphic properties of an encryption scheme we have to encrypt each input in its own cipher text. Following table shows the research in this field.

Table 3.1: homomorphic schemes[24]

Goldwasser-Micali, 1984[23]	Because of its restrictive message development during encryption (i.e. every 1 bit plaintext is encrypt as ciphertext of every 1024 bits)
Benaloh cryptosystem[9]	permits the encryption of bigger piece sizes at once
Paillier cryptosystem[11]	encodes 1024-bit messages in cipertexts of no less than 2048 bits, which is feasible on the off chance that we take include huge plaintexts.

Secret sharing was introduced independently by Shamir [7]. many other secret sharing schemes like Diffie hallman was also available in cryptography field but allows you to share secret between two parties with a prior establishment. security is dependent on security parameter used in scheme. In shamir's secret sharing schemes more number of parties can be involved and throshold t is decided to construct secret. Less then t parties trying to reconstruct secret will not be possible and they will fail to learn anything.

In the multi-party scenario[20], there are protocols that enable the parties to compute any joint function of their inputs without revealing any other information about the data which is given as input. That is, applying the function while attaining the same privacy as in the ideal model.

3.2 Privacy Preserving Distributed Data Mining(PPDDM)

Distribute data mining permits distributed sites owning individual data sets to perform mining by joining their data. Data is scattered to different sites. By applying mining algorithm locally, Local result model is prepared. Results are then aggregated using either trusted third party or secret sharing schemes. In distributed scenario, many problem arises while processing data for mining, like different naming conventions for attributes at different sites, arbitrarily partitioned data across sites. There are many real world problems in distributed databases where privacy of data is major concern. first, Different hospitals wants to mine their databases for research purpose, puts privacy of their patient in danger. Second, different intelligent agencies wants to mine their data without revealing any information about agencies or their operations. Due to this problems, different organization cannot directly share or pool their databases without preserving privacy. and ppddm aims to achieve this. For further understanding, following table shows the structure of algorithms used in privacy preserving distribute data mining.

Table 3.2: Distributed Data mining models for privacy preservation[24]

	PPDM algorithm
Data distribution	Horizontally, Vertically
Data mining Algorithm	ARM, Classification, Clustering
Communication model	Semi trusted third party, Collaborative computation based
Cryptography based technique	Oblivious transfer, homomorphic encryption, secret sharing based

3.2.1 Data Distribution Model

First we need to discover how the data are partitioned while applying PPDM algorithm. The relational databases are the most commonly used database in distributed scenario. therefor we focus on different data partitioned model in context of the relational model. In horizontal partitioned[24] dataset, different site collects same types of columns about the different databases. for example two organization collects same type of database. however customer database for each database schema might be different. This database structure usually occurs in same organization or across similar domains. For example two medical institutes viz. PS medical college and LJ hospital, each of which collects information

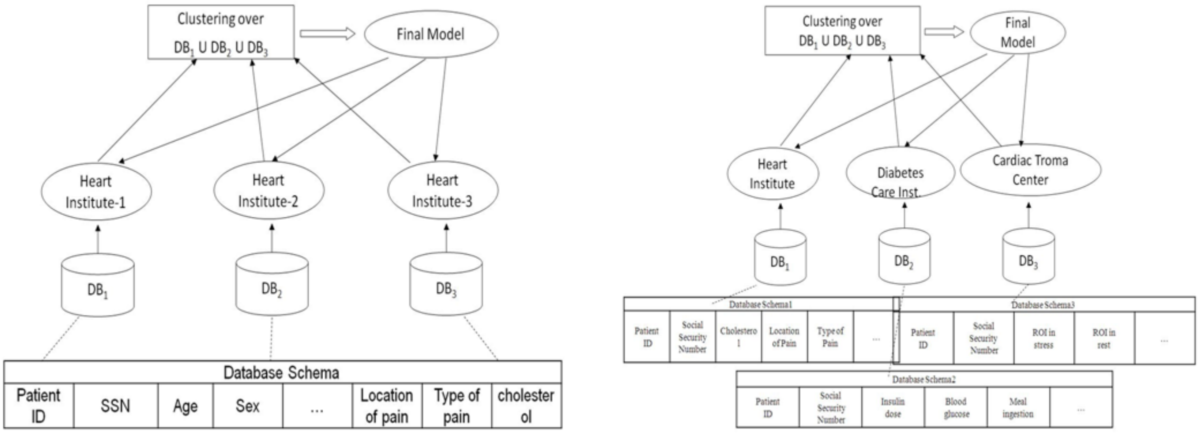

Figure 3.2: Horizontally and Vertically partitioned database

of their patients. attributes like patientcode, gender, occupation, age and disease are stored in both datasets. Merging two datasets gives more accurate predictive models. In vertically partitioned data, all organization have the same objects, but different types of attribute. for example database might be following type: in some city big bazar took their customers' information buying tomatoes and potatoes. in the same city, d mart(competitor of Bigbazar) gathers information of customers' buying beaf, onions. Now this two dataset have some linking information which helps in joining them. Mining of these two datasets outputs the buying behavior of customers' in that particular city. For vertically apportioned database it is expected that all organization have the same attributes, and that no organizations offer variables. With a specific end goal to match up a vertically partitioned[24] database, all organizations must have a worldwide identifier, for example, security number by government. Both the models are themselves useful in different scenarios. In some cases arbitrarily portioned data sets are used which consist of very complex structure to handle.

3.2.2 Oblivious transfer

In cryptography, a oblivious transfer protocol (regularly condensed OT) is a kind of protocol in which a sender exchanges one of possibly numerous bits of data to a collector, however stays secret regarding what piece (if any) has been exchanged. The first type of oblivious transfer[36] was presented in 1981 by Michael O. Rabin. In this structure, the sender makes an impression on the recipient with probability 1/2, while the sender stays oblivious in the matter of whether the collector got the message. Rabin's oblivious transfer scheme is focused around the RSA cryptosystem. A more valuable type of oblivious transfer called 1-2 oblivious transfer was produced later by Shimon Even, Oded Goldreich, and Abraham Lempel2, so as to construct conventions for secure multiparty processing. It is summed up to ”1 out of n oblivious transfer” where the client gets precisely one database component without the server getting to know which component was questioned, and without the client knowing anything about alternate components that were not recovered. The recent thought of oblivious transfer is a fortifying of private data recovery, in which the database is not kept private.

3.2.3 Homomorphic encryption

Homomorphic encryption system allows specific type of operations on encrypted data and outputs the results on encrypted data. If one wants to carry out operations like addition and multiplication on statistical encrypted data, there exist public key cryptography system for certain operations. The result is also an encrypted data. There are several partial homomorphic encryption system exist, which are less secure. fully homomorphic encryption systems are more secure and provide secure computation on encrypted data. In recent development in the field of cryptography, more secure encryption systems are developed for processing encrypted data and producing results but they have higher computation cost due to complexity of algorithms.

3.2.4 Secret Sharing based

secret sharing or secret splitting amongst N number of parties proposed by shamir. secret sharing schemes allow N number of parties to share secret without ant prior establishment of keys. If secret threshold is kept t, then minimum t number of parties have to gather to compute or to know about the whole secret. shamir’s secret sharing system shares secret using polynomials and random values with arbitrarily selected bits and then construct it back using lagrange’s interpolation equation.

Chapter 4

New Approach for PPDM: Functional Encryption(FE)

We have started new study in the field of cryptography called Functional Encryption[40]. It generates restricted keys to learn specific output of function on encrypted data, but learn nothing about the data. Encryption of data usually deals with securely sharing data over non secure network or data storage. Earlier in cryptography, if two parties want to communicate over non-secure channel, need to do a prior establishment to encrypt their data for non-secure line. While this is acceptable for two party case but large number parties generate larger cost of communication and computation. Nearly thirty years ago Diffie and hellman gave solution to share secret key without an a prior establishment or sharing any secret. Now it is time to study advanced topic in cryptography named Functional Encryption. In functional encryption [40] system, a decryption key allows user to learn a function of encrypted data. More precisely, in functional encryption system for function F (. ; .), a Third party holding MSK generates subkey sk that allow to compute function F (. ; .) on data which is encrypted. Dan Boneh and Amit sahai gave a brief definition and security proofs of functional encryption system. They gave simulated definition of FE. Same author has also defined data privacy and function privacy in functional encryption. Definition 1. We define Functional encryption[46] (FE) for function F over (K, X) is a tuple of four PPT(probabilistic polynomial time) algorithm following manner: (setup, keygen, encr, decr) satisfying the following:

- SETUP(SP) is a p.p.t algorithm that takes input as security parameter and outputs master public key and master secret key (MPK, MSK).

$$G_1,\ G_2\ of\ prime\ order\ p$$

$$G_1\ is\ addtivie\ notation$$

$$G_2\ is\ multiplivative\ notation$$

$$P,\ Q\ generators\ of\ G_1,\ we\ write$$

$$aP = p + p + p + p....P \rightarrow atimes$$

$$mappinge : G_1 \times G_2 \longrightarrow G_2$$

$$Bi - linear\ group\ G\ of\ prime\ order\ p,\ Generator\ g$$

$$Random\ exponent\ \alpha \in Z_p$$

$$Randomcomponents\ H_i \in G\ for\ each\ i \in U$$

$$PP := g, e(g,g)_{\alpha}, Hi \in U$$

$$MSK := \alpha$$

- KEYGEN(MSK, C) is a p.p.t algorithm that takes input master secret key MSK and circuit and Outputs corresponding secret key Sk.

$$split\ \alpha\ into\ shares\ \lambda_i\ following\ f$$

$$choose\ random\ r_i\ \ \in\ \ Z_p$$

$$SK = g^{\lambda_i}, H_i^{r_i}, g^{r_i}$$

- ENCRYPT(PK, x) is p.p.t algorithm that takes input as master public key PK and an input message x and outputs ciphertext CT.

$$choose\ random\ s \in Z_p$$

$$CT\ \ =\ \ Me(g,g)^{\alpha s}, g^s, H_{i\ i\in S}^{s}$$

- DECRYPT(SK, CT) is deterministic algorithm that takes input as secret key SK, CT and outputs C(x).

$$CT = \quad g^s \qquad H_i^s$$

$$SK = \quad g^{\lambda_i} H_i^{r_i} \qquad g^{r_i}$$

$$e(g,g)^{\lambda_i s} e(g,H_i)^{r_i s} \qquad e(g,H_i)^{r_i s}$$

$$we\ eliminate\ randomness\ at\ time\ of\ decryption$$

$$and\ through\ enough\ share\ we\ can$$

$$reconstruct\ \alpha\ in\ exponent$$

4.1 Subclass of functional encryption

For applicability of functional encryption following two classes are defined

- Predicate encryption with index.
- Predicate encryption without index.

4.2 Predicate encryption with index

we start our study with simplest encryption case of Identity based encryption[46] in functional encryption and move towards advance path with attribute based encryption.

4.2.1 Identity based encryption

In Identity based encryption ciphertrxt and secret keys are connected with characters and a secretkey can decrypt a ciphertext if two of them are equivalent. IBE shows the first utility which isn't straightforwardly feasible from asymmetric key encryption. Boneh and Franklin and Cocks build first practical development of IBE system, where demonstrated secure as indicated by indistinguishable definition. Another schemes were demonstrated secure under the standard oracle model but under selective security and adaptively secure.

4.2.2 Attribute based encryption

sahai and Waters defined ABE where complex access policies are expressed. Then Goyal, Pandey, Sahai[41]and Waters formulate two different ABE schemes. KEY policy ABE, Ciphertext policy ABE. In KP abe, attributes are attached with key which are distributed for decryption of cipher text for allowing decryption to only those data owned by particular party. In cipher text policy based abe, policy for allowing decryption or learning functions are attached to cipher text. In both the cases, if the policy is satisfied, then only party is allowed to decrypt ciphertext.

4.3 Predicate encryption without index

Predicate without public index is useful when faster results are needed without publicizing index associated with data. while above scheme take into account expressive types of accessing mechanism, they are restricted in following points. Firstly, the index associated is a part of the empty functionality which is given clearly, this itself is private information. Secondly, computation on data which is encrypted is not allowed. following is predicate encryption system that don't leak index.

4.3.1 Inner product based encryption

Katz, Sahai and Waters[41] proposed framework for testing if dot operstion over the ring Zn is equivalent to 0, where N is a result of three arbitrary prime picked by the setup PPT algorithm. Inner product operation proven to be faster approach in ceratin application of functional encryption. In this approach vactors of key, vector of random value and vector of data have to be at 90 degree to each other for dot product to be 0. After that Okamotoand and Takshima gave development over the field Fp.

Chapter 5

Proposed Framework and Implementation

Two framework to incorporate functional encryption in privacy preservation for data mining application have been proposed.

- Semi Trusted Third Party(STTP) model with centralized approach
- Semi Trusted Authority(STA) with collaborative processing

We have incorporated Identity Based Functional Encryption scheme of sahai and waters in STTP model. In this scheme, users in data mining process are provided unique Identifier. Due to the property of non collusion of functional encryption, users can't collude for results. In STTP model, key distribution, result aggregation, final result distribution and all central authority related work is done by STTP. If STTP becomes malicious user in this model, then also there are very rare chances that it gets control over final results of data mining due to functional encryption properties. In Semi Trusted Authority model, only key distribution and key generation phase in handled by STA. STA model is collaborative processing model where each user has to input their data and get the result according to its input. STA model exhibits more communication cost than STTP due to more number of communication round. Computation cost is also higher in collaborative proccesing model.

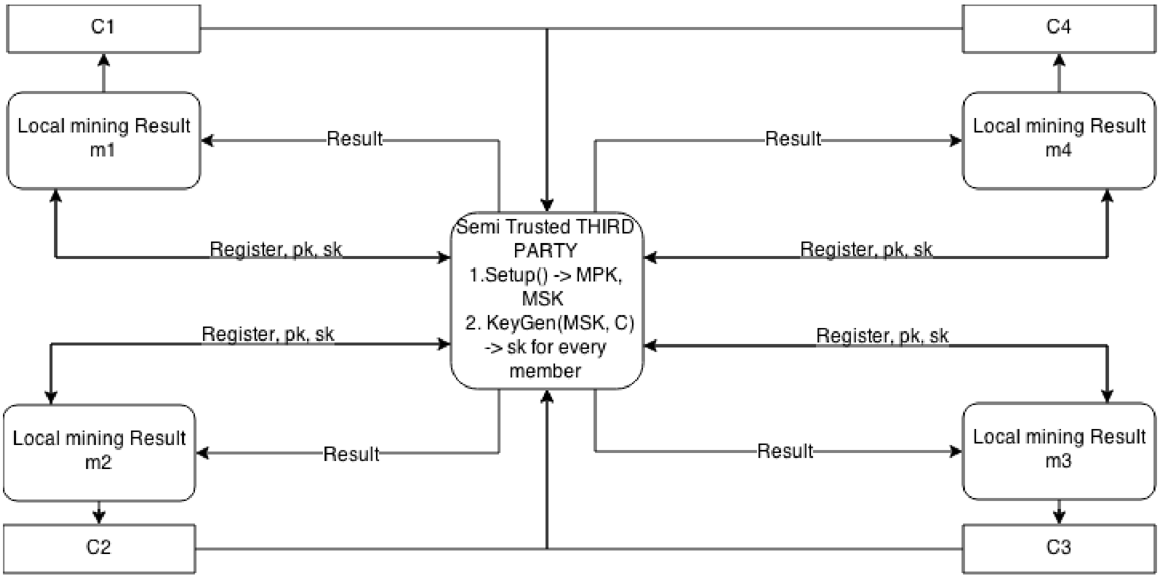

Figure 5.1: Semi Trusted Authority Model

For STTP model, this is referred bootstrapping theorem of Sergey, Vindo and Hoeteck which uses homomorphic encryption scheme incorporating functional encryption Theorem states that Let BDFE be a Q question FE plan which meets expectations for any circuit NC1 and HE be a symmantically secure homomorphic encryption conspire whose unscrambling calculation $HE.Dec(SK, ct)$ can be actualized by NC1 circuit in the mystery key. At that point for any group of poly size circuits C there exist Q inquiry FE plan $FE = (FE.setup, FE.keygen, FE.enc, FE, dec)$ [43]. Our proposed algorithm in STA model works as follows: here we run two components of this scheme to make proper use of this theorem.

1. Inner circle encryption scheme
2. Outer circle encryption scheme.

Now STA starts process by running SETUP algorithm.

pseudo code:privacy preserving k mean clustering incorporating Functional encryption using Secure Trusted Authority

Setup $FE.setup(1^k)$ **:** This algorithm produce pair of master public and secret key

$$FE.SETUP(1_k) \rightarrow (MPK, MSK)$$

Key Generation $FE.keygen(MSK, C)$**:** It takes input to master secret key and circuit C, run key generation algorithm to generate key for circuit member

$$FE.keygen \rightarrow SK_c$$

Encryption $FE.enc(MPK, ct)$**:** This algorithm takes input as cipher text by circuit member, unique id of member and master public key, give output as encrypted text to STTP. before third step, inner circle FE scheme runs as follows

1. Produce uniformly random public and private key pair for homomorphic encryption scheme for individual parties

   $$HE.keygen(1^k) \rightarrow (PK, SK)$$

2. Encrypt input text using public key of homomorphic encryption scheme

   $$HE.Enc(PK, x) \rightarrow ct$$

   from this step, ct is send to STTP from all the party and processed for aggregation of results. Now assumption shows that STTP has tremendous power to perform aggregation at that time another cipher text of final result is generated. STTP run bounded FE.HE algorithm.

3. Run bounded FE algorithm to encrypt cipher text ct together with homomophic secret key

   $$FE.HE.enc(MPK, (ct, SK) \rightarrow CT$$

   output CT as cipher text.

Decryption $FE.dec(Sk_c, CT)$**:** This algorithm performs in following manner.

$$FE.HE.dec(SK_c, CT) \rightarrow HE.dec(ct, sk) \rightarrow x(finalresult)$$

---

so at individual party, when they enter their secret key with unique identifier and if it is accepted then in output they only learn final result but nothing else.
This paper have also proposed framework that enables parties to share secret in between and compute results securely. This framework contain Secure Trusted Authority model, Secure multiparty computation with shamir's secret sharing scheme, Identity based functional encryption scheme. IDFE scheme is newly introduced in our framework to increase the measure of privacy and security. Our framework exhibits higher cost against other research but it is stable and standard approach that can be accommodated to real world scenarios where sensitive information needs to be protected.

---

pseudo code:privacy preserving k mean clustering incorporating Functional encryption using Secure multiparty computation

---

Setup $FE.setup(1^k)$: This algorithm produce pair of master public and secret key

$$FE.SETUP(1_k) \rightarrow (MPK, MSK)$$

Key Generation $FE.keygen(MSK, C)$ It takes input to master secret key and circuit C, run key generation algorithm to generate key for circuit member

$$FE.keygen(MSK, C) \rightarrow SK_c$$

At this time each party would have received their pair of keys.Now **Distributed K mean** algorithm starts.
n number of parties

```
k number of clusters
M number of attributes
$\mu_i$ number of objects
$\mu_1.....\mu_k$ Initialize k number of cluster at each party i
Do parallel for each party
calculate $x_i$ object nearest to $\mu$
for j := 1 to k step 1 do // let $\mu_{ji}$ be the j-th cluster to party i
compute for each party for each cluster
for m = 1 to M, for each party
Encrypt( $\mu_{ij}^{m}, m \subset U$ )
end for
end for
for m= 1 to m
call secretsharing( $\mu_{ij}^{m}, n$ )
end for
call secretsharing( $mu_{ij}^{m}, \mu$ )
compute sum of shares and call it $S(x_i)$
send and receive $S(x_i)$ from all parties
Solve the set of equation using lagrange's interpolation to find the sum of secret values
Compute global cluster mean for j th cluster
end for
recompute $\mu_i$ using global cluster mean until termination criteria met
return $\mu_i...\mu_k$
for m=1 to M for each party
Decrypt($sk_i, m, \mu_i$ )
end for
```

STA model of privacy preservation is more costlier than earlier approach. Each party in this approach is assumed to be semi honest to perform operations for data mining. We have tried to include more security by adding extra encryption cost which enhances security of our model.

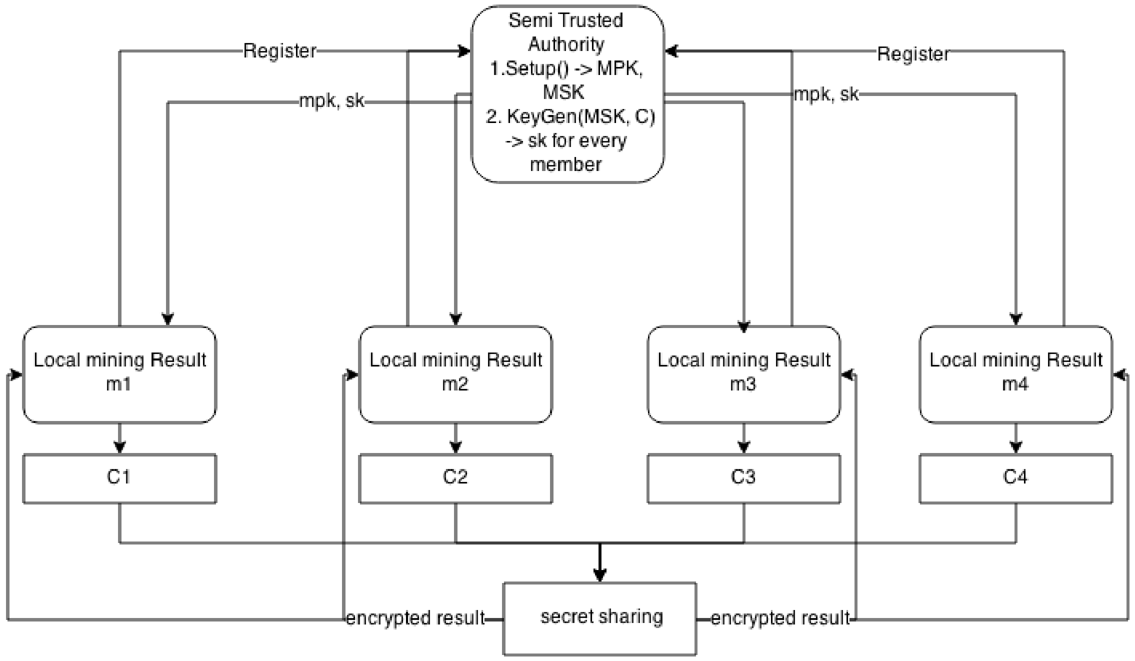

Figure 5.2: Semi Trusted Authority Model

Chapter 6

Implementation and Analysis

6.1 Performance Analysis

This study has proposed two different architectures based on functional encryption schemes. As functional encryption is most advanced approach in cryptography, it is based on many assumptions and limitation. study has provided balanced architectures with security and privacy provisions. we have analyzed our approach on security provisions and also with cost estimation of models. Though results tend to be higher but our approach provides more security and privacy measures.

6.1.1 Security Analysis

Our approach used Identity based functional encryption scheme and Attribute based FE which are based on non adaptive, bounded approach. In non adaptive environment, this approach provides selective security. This limitation of FE binds user to behave in certain manner to get the key pairs. FE links computation together to bind user to follow certain steps to get key pair. Usage of bi linear groups provides cryptography mapping between the keys pairs which uses elliptic curves. Non degeneracy of bi linear group exhibits following property:

$$p \neq q \Rightarrow e(p,p) \neq 1$$

FE exhibits two types of definition SIM-based and IND-based for security provisions. We have used SIM-based definition which secure against selectively secure environment in non adaptive approach. For security provisions, FE uses linear secret sharing schemes which makes cost of security lower and provides higher security measures. Independent

Randomness also makes it more easier at the end for final results due to its cancellation from cipher text.

6.1.2 Cost Analysis

Functional encryption uses bi linear groups. Bi linear groups have properties like bi linearity and computationally efficient. bi linearity

$$\forall\ P,\ Q \in\ G,\ \forall\ a,\ b\ \in\ Z_q$$

$$e(aP, bQ) \Rightarrow e(P, Q)^{ab}$$

and e computationally efficient is because G_1 is elliptic curve group and G_2 is finite group. Proposed approach uses Functional encryption with Secure multiparty computation with Shamir's secret sharing scheme which means cost goes higher than earlier research in same area. Decryption complexity of FE scheme is $\sqrt{m}$. Size of this groups is small. They contain more complex calculation and they more efficient then any other groups. With the increasing cost of calculation, we are also providing standardized way to use privacy preserving data mining model with higher provision of security and privacy. We have also maintained balance between these two factors.
Ubuntu/linux with 1GB RAM and 60 GB hard disk machine was used for implementation.Setup of experiments contain 2 machine in same lan. Our assumption was that data is horizontally partitioned on both machine equally. Both machine contain same amount of data. Several experiments for future use have been conducted. First K means clustering on Standard databases from UCI repository IRISH, DIABETES, ELECTRICITY BOARD with different sizes was implemented in python .Paillier homomorphic encryption system with key size 512, 1024, 2048 was implemented.Charm-Crypto Framework and implemented Attributed based functional encryption scheme proposed by Sahai and Waters have been used.

Results shown above are expensive in terms of computation. Analysis shows that with increasing number of parties, cost may go higher than expected. higher number of objects in database exhibits higher cost.

Table 6.1: K means Algorithm Analysis

Criteria	K-means Running Time	K-means Accuracy	Number of Attributes	Number of Records
IRIS	0.03	88.667	05	150
DIABETES	0.06	51.6927	09	768
ELECTRICITY BOARD	3.12	76.98	05	45781

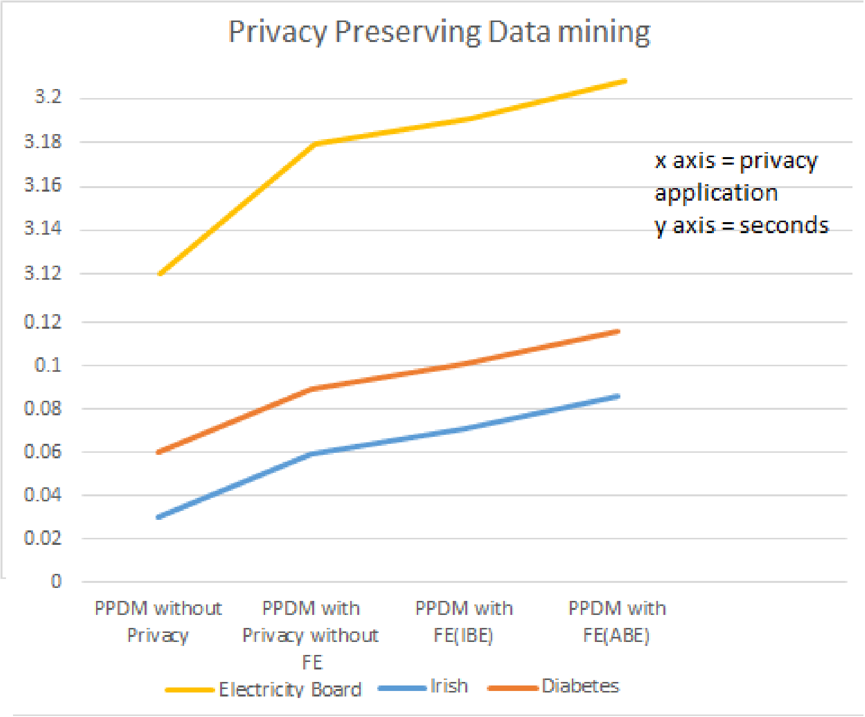

Figure 6.1: Computation cost comparison with earlier approach

Table 6.2: Computation Cost analysis

seconds(s)	IRISH	DIABETES	ELECTRICITY BOARD
Without Privacy	0.03	0.06	3.12
PPDM with privacy without FE	0.059	0.089	3.179
PPDM with privacy with FE(IBE)	0.071	0.101	3.191
PPDM with privacy with FE(ABE)	0.086	0.116	3.208

Chapter 7

Conclusion and Future work

This master thesis has tried to include most of technique to survey on privacy preserving data mining. Till now researches is working on standardize model for privacy preserving data mining. Big data has come into picture and privacy concerns have grown more for data to be secure for individual privacy. Privacy preservation can be applied to certain limit based on data mining algorithm. Privacy and accuracy is a pair of contradiction. One overrides the result of another. In distributed environment, we concern about finding balance between algorithm complexities, computational cost and security. Number of data mining algorithms are proposed so far, but not a single algorithm is up to the marks. Proposed architectures of ppddm exhibits higher amount of cost. Against that, they provide higher measure of privacy and security. They are secure against security attacks like CCA2, CPA. They also secure against privacy attacks like inconsistent shares with honest party, consistent shares with adversarial party. Proposed architecture still is based on many cryptography assumptions. In future, more number of parties will be included in experiment, large amount of data will be included. Proposed architecture needs to be tested under real world scenarios

References

[1] Ontario. Information and Privacy Commissioner, and Ann Cavoukian. Data mining: Staking a claim on your privacy. 1997.

[2] Liu Yu, Dap eng L, et al, Survey of research on anonymilization technology in data publication, Computer Application, pp. 2361-2364, 2009

[3] Verykios, Vassilios S., et al. "State-of-the-art in privacy preserving data mining." ACM Sigmod Record 33.1 (2004): 50-57.

[4] R. Agrawal and S. Ramakrishanan, Privacy preserving data mining. ACM sigmod record, 2004.

[5] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao-Kui. Privacy Preservation in Database Applications: A Survey. Chinese journerl of computer,2009

[6] Yan Zhao1 Ming Du2 Jiajin, Le1 Yongcheng Luo1, A Survey on Privacy Preserving Approaches in Data Publishing. First International Workshop on Database Technology and Applications, 2009

[7] A. Shamir. How to share a secret. Communications of the ACM, 22(11):612 613, November 1979.

[8] A. C. Yao. Protocols for secure computations (extended abstract). In 23rd Annual Symposium on Foundations of Computer Science. IEEE, 1982.

[9] J. Benaloh. Dense probabilistic encryption. citeseer.ist.psu.edu/benaloh94dense.html, 1994.

[10] Oliveira, Stanley RM, and Osmar R. Zaiane. "Privacy preserving frequent itemset mining." Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14. Australian Computer Society, Inc., 2002.

[11] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In Advances in Cryptology EUROCRYPT'99, pages 223-238. Springer, 1999.

[12] xiaolin Z. and Hongjing B. Research on privacy preserving classification data mining based on random perturbation. National conference of Information. Vol 1. No 1.,2010.

[13] Kamakhi P. and Vinnaiya babu. Preserving privacy and sharing the data using classification on perturbed data. IJSCE. Vol 2. No 3. 2010.

[14] Jagannathan, Geetha, and Rebecca N. Wright. "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

[15] Kantarcioglu, Murat, and Chris Clifton. "Privacy-preserving distributed mining of association rules on horizontally partitioned data." IEEE Transactions on Knowledge and Data Engineering 16.9 (2004): 1026-1037.

[16] Patel, Sankita, Sweta Garasia, and Devesh Jinwala. "An Efficient Approach for Privacy Preserving Distributed K-Means Clustering Based on Shamirs Secret Sharing Scheme." Trust Management VI. Springer Berlin Heidelberg, 2012.

[17] Kantarcoglu, Murat, Jaideep Vaidya, and C. Clifton. "Privacy preserving naive bayes classifier for horizontally partitioned data." IEEE ICDM workshop on privacy preserving data mining. 2003.

[18] Vaidya, Jaideep, and Chris Clifton. "Privacy-preserving k-means clustering over vertically partitioned data." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[19] Samarati, Pierangela. "Protecting respondents identities in microdata release." Knowledge and Data Engineering, IEEE Transactions on 13.6 (2001): 1010-1027.

[20] Duan, Yitao, and John F. Canny. "Practical Private Computation and Zero-Knowledge Tools for Privacy-Preserving Distributed Data Mining." SDM. 2008.

[21] Friedman, Arik, Assaf Schuster, and Ran Wolff. "k-Anonymous decision tree induction." Knowledge Discovery in Databases: PKDD 2006. Springer Berlin Heidelberg, 2006. 151-162.

[22] Blanton, Marina. "Achieving full security in privacy-preserving data mining."Privacy, security, risk and trust (passat), 2011 ieee third international conference on social computing (socialcom). IEEE, 2011.

[23] Yang, Bin, et al. "Collusion-resistant privacy-preserving data mining."Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[24] Xu, Zhuojia, and Xun Yi. "Classification of privacy-preserving distributed data mining protocols." Digital Information Management (ICDIM), 2011 Sixth International Conference on. IEEE, 2011.

[25] Fung, Benjamin CM, Ke Wang, and Philip S. Yu. "Anonymizing classification data for privacy preservation.Knowledge and Data Engineering, IEEE Transactions on 19.5 (2007): 711-725.

[26] Fang, Weiwei, and Bingru Yang. "Privacy preserving decision tree learning over vertically partitioned data." Computer Science and Software Engineering, 2008 International Conference on. Vol. 3. IEEE, 2008.

[27] Dasseni, Elena, et al. "Hiding association rules by using confidence and support." Information Hiding. Springer Berlin Heidelberg, 2001.

[28] Lin, Zhenmin, and Jerzy W. Jaromczyk. "Privacy preserving two-party k-means clustering over vertically partitioned dataset." Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on. IEEE, 2011.

[29] Slavkovic, Aleksandra B., Yuval Nardi, and Matthew M. Tibbits. "" Secure" Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases." Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. IEEE, 2007.

[30] Xiao, Ming-Jun, et al. "Privacy preserving id3 algorithm over horizontally partitioned data." Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on. IEEE, 2005.

[31] Xiao, Ming-Jun, et al. ”Privacy preserving C4. 5 algorithm over horizontally partitioned data.” Grid and Cooperative Computing, 2006. GCC 2006. Fifth International Conference. IEEE, 2006.

[32] Inan, Ali, et al. ”Privacy preserving clustering on horizontally partitioned data Data and Knowledge Engineering 63.3 (2007): 646-666.

[33] Pang, Liaojun, et al. ”A verifiable (t, n) multiple secret sharing scheme and its analyses.” Electronic Commerce and Security, 2008 International Symposium on. IEEE, 2008.

[34] Aggarwal, Charu C., and S. Yu Philip. ”A condensation approach to privacy preserving data mining.” Advances in Database Technology-EDBT 2004. Springer Berlin Heidelberg, 2004. 183-199.

[35] Reza, M., and Somayyeh Seifi. ”Classification and Evaluation the PPDM Techniques by using a data Modification-based framework.” IJCSE, Vol3. No2 Feb (2011).

[36] Pinkas, Benny. ”Cryptographic techniques for privacy-preserving data mining ACM SIGKDD Explorations Newsletter 4.2 (2002).

[37] Pedersen, Thomas Brochmann, Ycel Saygn, and Erkay Sava. ”Secret charing vs. encryption-based techniques for privacy preserving data mining.” (2007)

[38] Taylor, Ronald C. ”An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics.” BMC bioinformatics 11.Suppl 12 (2010): S1.

[39] Naveed, Muhammad, et al. ”Controlled Functional Encryption.” Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, November-2014. Beimel, Amos, et al. ”Non-Interactive Secure Multiparty Computation.”Advances in CryptologyCRYPTO 2014. Springer Berlin Heidelberg, 2014. 387-404.

[40] Agrawal, Shashank, et al. ”Function Private Functional Encryption and Property Preserving Encryption: New Definitions and Positive Results.” IACR Cryptology ePrint Archive 2013 (2013): 744

[41] Attrapadung, Nuttapong, and Benot Libert. ”Functional encryption for public-attribute inner products: Achieving constant-size ciphertexts with adaptive security or support for negation.” J. Mathematical Cryptology 5.2 (2012): 115-158.

[42] Barbosa, Manuel, and Pooya Farshim. ”Delegatable homomorphic encryption with applications to secure outsourcing of computation.” Topics in CryptologyCT-RSA 2012. Springer Berlin Heidelberg, 2012. 296-312.

[43] Gorbunov, Sergey, Vinod Vaikuntanathan, and Hoeteck Wee. ”Functional encryption with bounded collusions via multi-party computation.” Advances in Cryptology-CRYPTO 2012. Springer Berlin Heidelberg, 2012. 162-179.

[44] Boneh, Dan, Amit Sahai, and Brent Waters. ”Functional encryption: Definitions and challenges.” Theory of Cryptography. Springer Berlin Heidelberg, 2011. 253-273.

[45] Yang, Xiaoyuan, Weiyi Cai, and Ping Wei. ”Multiple-authority-keys CP-ABE.”Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on. IEEE, 2011.

[46] Lewko, Allison, et al. ”Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption.” Advances in CryptologyEUROCRYPT 2010. Springer Berlin Heidelberg, 2010. 62-91.