Secure Multi Party Computation Technique for Classification Rule Sharing

Murugeshwari B
Department of Computer Science & Engg
S.A. Engg College
Chennai,Tamilnadu,
India.

Jayakumar C
Department of Computer Science & Engg
R.M.K Engg College
Chennai,Tamilnadu,
India

Sarukesi K
Hindustan University
Chennai, Tamilnadu
India

ABSTRACT

Confidentiality of data or resources is of primary importance in Privacy Preserving Data Mining (PPDM) Systems. The research work presented through this paper discusses the PPDM model in which the privacy of data transacted amongst the various Data Custodians involved is highlighted. The data available with each data custodian is assumed to be horizontally portioned. The proposed model considers the C5.0 algorithm for data mining and classification rule generation due to its advances and classification accuracy over its predecessors. Privacy of the data transacted or secure multiparty computation is achieved by using the commutative RSA cryptography scheme. The proposed model is compared with the existing secure group communication techniques like Secure Lock and Asynchronous Control Polynomial in terms of computational efficiency. Furthermore the privacy preserving feature of the proposed scheme is proved in terms of the computational indistinguishablity of the data transacted amongst the varied data custodians involved discussed in the paper.

Keywords
Privacy Preserving Data Mining; Semi Honest Model; Secure Multiparty Computation; Commutative RSA; C5.0 Data mining Algorithm; Classification Rules;

1. INTRODUCTION

Business Corporations, Government Departments, Research organizations, profit and non-profit organizations, attain data and analyze the data to fulfil their respective desired goals. The data utilized generally embodies classified information of private entities. Confidentiality is a major security risk associated with such data especially when it is available electronically. Securing classified statistical databases has always drawn the attention of researchers and government, non-government organizations [1],[2],[3],[4]. To address the issue of security and confidentiality governing bodies like European Union Privacy Directive, US Health Insurance Portability and Accountability etc have incorporated imposing mandatory norms to provide for the security and confidentiality of data and its use. The statistical data discussed here is utilized by research communities, business corporations, health care organizations, and government departments to derive knowledge and understand the trends exhibited by the data using data mining techniques. The accuracy of the data mining technique adopted depends on the correctness of the data used for the purpose of mining. The research work presented here assumes that the data is distributed amongst multiple data custodians which is the case most often than not. For this reason techniques utilized for the purpose for mining are viewed as threat to data confidentiality [5][6][7]. For example considering the domain of healthcare data is available with multiple sites pertaining to a particular disease. The sites housing the data contain confidential data about patients which cannot be disclosed or shared. To address the issue researchers have proposed various Privacy Preserving Data Mining (PPDM) techniques amongst the multiple data custodians involved. An ideal PPDM system incorporates privacy preservation of confidential data yet preserving the effectiveness of the data mining outcome. Multiple PPDM techniques that currently exist address this issue by incorporating algorithmic approaches and mathematical tools rather than concentrating on the system design[8][9][10]. The research work presented in this paper primarily discusses a systemic view and the design concepts incorporated to address the issue of privacy preserving.

Classification of data is considered as a vital technique for data mining. A Classifier (trained using a classification algorithm) could be utilized to classify unseen data without the need for data sharing amongst the custodians using classification rules. The approach discussed in this paper discusses the use of rules sharing for classification rather than the data utilized to generate the rules, securing privacy of the data and also achieving effective data mining results.

The remaining manuscript is structured as follows: Section 2 discusses the related work of privacy preserving data mining systems in brief. The next section introduces the proposed system model construction for privacy preserving in 6 phases. Section 4 provides the comparisons of the proposed model with existing group communication systems. The penultimate section of the paper discusses the evaluation results proving the proposed model’s efficiency to provide privacy of the data exchanged amongst the data custodians in terms of computational indistinguishablity. The conclusion and the future scope of the research work are discussed in Section 6 of this paper. The research work presented here is aimed to address the issue of privacy preservation in a PPDM system eliminating the need for any key exchange to facilitate secure communication amongst the data custodians and securing the data.

2. RELATED WORK-PRIVACY PRESERVING DATA MINING

PPDM was introduced initially in the year 2000 [11][12] in which many issues related to this active area of research were discussed. Researchers from then on have proposed varied models to address the issues related to PPDM. PPDM could be broadly classified into two categories, derived based on the privacy level provided to the data held by each party or data custodians as in referred here namely Secure Multiparty computation and partial information hiding [13].The first category or Secure Multiparty Computation provides a robust level of security where as the other category of partial information hiding provides lower levels of privacy and improved mining performances. Considerable amount of work has been carried out in the area of Secure Multiparty computation [14] [15]. The data considered with each data custodian could be horizontally portioned [11] [16][27], vertically portioned [17] [18] or even clustered using k-means algorithm [19]. Research work utilizing these models have been closely studied prior to the development of the proposed model. The arena of PPDM using information hiding could be classified into data perturbation [12][20][21][22] , retention replacement [23] [24][25] and k-anoymity [28][29][30]. The proposed system considers secure multiparty computation to provide for more privacy and also considers the C5.0 [40] algorithm to provide for better mining results when compared to its predecessors like ID3 [19][40] and C4.5 [16][18] decision tree algorithms.

To preserve the privacy of the data to be transacted researchers have proposed varied privacy preserving techniques like noise inclusion techniques[32][33], cryptographic techniques[12][13][16][17][33], anonomization of the data [28][29][3] and many more. Cryptographic approaches for securing are more preferred as they preserve the data integrity. The approach proposed in his paper also utilizes the advantages of cryptographic schemes to maintain data integrity. Secure group communications based on cryptographic techniques like Secure Lock [34] , Access Control Polynomial [35][36] have been studied for comparisons with our proposed scheme. Based on the literature survey conducted it was found that cryptographic approaches are generally used to provide for data security but the cryptographic techniques utilized involve numerous overheads and criticalities like key distribution , rekeying mechanisms ,key computations to provide for security [37][38][39].The scheme proposed in this paper is an attempt to reduce the overheads involved in key management and distribution by utilizing the benefits of symmetric cryptographic techniques with no key exchange.

3. PROPOSED PRIVACY PRESERVING DATA MINING SYSTEM

3.1 Preliminary notations

Let us consider a training database $\mathcal{D}_m$ to be considered for mining. Let a set $\mathcal{C} = \{c_1, c_2, c_3 \dots\dots . c_p\}$ represent the set of data custodians amongst which the database $\mathcal{D}_m$ is portioned horizontally. Also p represents the total number of data custodians. Let the data available with each custodian be represented by $d_{c\,x}$. Where $c \in \mathcal{C}$ and $1 \le x \le p$. Then the database $\mathcal{D}_m$ could be represented as

$\mathcal{D}_m = \{ d_{c\,1} \cup d_{c\,2} \cup d_{c\,3} \cup \dots\dots\dots d_{c\,p} \}$

The data available with the custodian's $d_{c\,x}$ consists of a set of transactions T defined as.

$T = \{ t_1 , t_2 , t_3 , \dots\dots . t_a \}$ Where a represents the total number of transactions

Each Transaction consists of items represented by the items set I represented as

$I = \{ i_1 , i_2 , \dots\dots\dots\dots i_b \}$

Where b represents the total number of attributes present per transaction .

The data $d_{c\,x}$ consists of a set of transactions T such that $T \subseteq I$ .Let us consider an item set $L \subseteq I$ and an transaction T contains L if and only if $L \subseteq T$.

The research work presented here considers the C5.0 Algorithm for data mining. Classification of data available with various custodians based on rules obtained from each custodian is considered in the proposed approach. A classification rule could be represented as $A \rightarrow B$. Consider the rule $A \rightarrow B$ has a $coverage\ C_o$ in the database $d_{c\,x}$ if C_o% of the transactions T in $d_{c\,x}$ contain $A \cup B$. The rule $A \rightarrow B$ holds in the database $d_{c\,x}$ with $confidence\ \ C_f$ if C_f % of T transactions in $d_{c\,x}$ that contain both $A\ and\ B$.

It is assumed that each data custodian embodies two datasets one training dataset $d_{c\,x}$ which is pre classified and another dataset $dt_{c\,x}$ which represents a test dataset or an unclassified dataset. The goal of the research work proposed here is to mine the test dataset $dt_{c\,x}$ available with each data custodian in a semi-honest model, securely without disclosure of any data $dt_{c\,x}$ or $d_{c\,x}$ amongst the varied custodians involved.

3.2 Proposed Privacy Preserving Data Mining Method

The privacy preserving data mining model proposed in this paper follows a 6 phase approach. The proposed model does not incorporate any data base exchanges $d_{c\,x}$ maintained by the custodians but rather relies on the locally generated rule exchange over secure channels to facilitate mining. The model described here is secure as no custodian discloses any data (partial/masked or complete) available with him maintaining privacy of data one of the primary goals of the research work presented here. The C5.0 algorithm was selected for the purpose of mining for its advantages over its predecessors like ID3[19][40] and C4.5[16][18].

The research work discussed in this paper considers a semi honest model of secure multiparty computation technique . The first phase uses the C5.0 data mining algorithm to generate the classification rules applicable to the data maintained by the data custodian locally. The data is considered to be horizontally portioned and available with $\mathcal{C}_p$ data custodians. Considering a semi honest trust model the next phase incorporates establishment of the commutative RSA algorithm amongst the $\mathcal{C}_p$ data custodians. In the second phase the encryption and decryption keys to be used for cryptographic based secure transfers are initialized. The rules that are generated locally in phase 1 are encrypted using the encryption keys of phase two in phase 3. All the $\mathcal{C}_p$ data custodians encrypt their respective classification rules. The encrypted rules of the data custodians are shared amongst themselves resulting in the construction of the combined secure rule set in phase 4. In penultimate phase each data custodian cumulatively decrypt the secure rule set to form the combined rule set. In the last phase the data custodian uses the Combined Rule Set to mine their data.

Phase 1: Using C5.0 generate local classification rules
Phase 2: Establish Commutative RSA agreement amongst custodians
Phase 3: Secure rule sharing amongst parties
Phase 4: Construction of Combined Secure Rule Set.
Phase 5: Construction of Combined Rule Set using commutative decryption techniques.
Phase 6: Using the Combined rule set to mine the data using C5.0

Phase 1: Using C5.0 Data Mining Engine to generate Classification Rules Locally

Let us consider $\mathcal{DE}_p$ represents the data mining engine available at each site. Where p represents the data custodian and $c_p \in C$. The data mining engine adopted in the proposed engine is based on the C5.0 algorithm. The engine considered for the initial phase utilizes the local data available with the data custodian to generate classification rules based on the pre decided coverage C_o and confidence C_f amongst the data custodians. Let R_{rp} represent the rules set generated at the p^{th} data custodian. The data mining engine $\mathcal{DE}_p$ could be represented as a function of the data $d_{c\,\mathcal{p}}$ available at the p^{th} data custodian who provides the initial rules R_{rp} .The data mining engine could be defined as

$\mathcal{DE}_p = f_r\left(d_{c\,\mathcal{p}}\,,\ C_o\,,\ C_f\right) = R_{rp}.$

Where the rule set R_{rp} is define by

$R_{rp} = \{\, r_{1p}\,, r_{2p}, r_{3p}, \dots \dots\ r_{np}\,\}$ Where n represents the total rules generated at p^{th} data custodian. The phase 1 of the proposed scheme could be represented as the following algorithm.

Input: Data $d_{c\,\mathcal{p}}$ available at the p^{th} data custodian, coverage C_o and confidence C_f.

Output: Locally generated rule set R_{rp}.

Algorithm 1:

```
1. For each data custodian $c_{\mathcal{p}}$ in the custodian set $\mathcal{C} = \{c_1, c_2, c_3 \dots \dots . c_p\}$
2.     Initialize pre agreed coverage $C_O$ and confidence $C_f$ amongst $\mathcal{C}$ and data $d_{c\,\mathcal{p}}$.
3.     Initialize $\mathcal{DE}_p = f_r\left(d_{c\,\mathcal{p}}\,,\ C_o\,,\ C_f\right)$
4.     Obtain $R_{rp}$ rule set generated
5. End For
```

Phase 2: Commutative RSA agreement amongst custodians.

Commutative cryptography is an important technique adopted in the proposed approach. The research work presented here utilizes the advantages of commutative encryption for secure exchange of rules generated by each data custodian and creation of the secure rule set. Let us consider D represents the data to be secured and K represent the Keys to used for encryption and decryption such that $K_1, K_2, \dots . K_l \in K$. An cryptographic algorithm is said to be commutative for any permutations of p, q if

$E_{K_{p1}}\left(\dots . E_{K_{pl}}(D) \dots \dots \right) = E_{K_{q1}}\left(\dots . E_{K_{ql}}(D) \dots \dots \right)$

And $\forall\, D_1, D_2 \in D$ where $D_1 \neq D_2$ for a given $k\,, \epsilon < {}^{1}\!/_{2^k}$

$\mathcal{Pr}\left(E_{K_{p1}}\left(\dots . E_{K_{ql}}(D_1) \dots \dots \right)\right) = E_{K_{q1}}\left(\dots . E_{K_{ql}}(D_2) \dots \dots \right) < \epsilon.$

Where $\mathcal{Pr}$ represents the probability and ϵ represents the probability factor.

With the above definitions it is evident that commutative cryptography could be used to check the data are equivalent without revealing any information. For example let us consider two custodians I and J having data D_i and D_j who would like to securely exchange their data using commutative cryptography. Custodian I sends its encrypted data represented by $E_{K_i}(D_i)$ to custodian. Likewise J sends it encrypted $E_{K_j}(D_j)$ data to I .Note that neither I or J can view the original data possesses by the other custodian. They can only view the data of the other custodian in an encrypted form. Now each custodian encrypts the received encrypted data $E_{K_i}(D_i)$, $E_{K_j}(D_j)$ using its own keys resulting in $E_{K_i}\left(E_{K_j}(D_j)\right)$ and $E_{K_j}\left(E_{K_i}(D_i)\right)$ available with custodian I and J respectively. Each custodian I and J can compare both the encrypted data $E_{K_i}\left(E_{K_j}(D_j)\right)$ and $E_{K_j}\left(E_{K_i}(D_i)\right)$. If the values obtained are varying it ensures secure transactions with high probability that $D_i \neq D_j$. If the values obtained are equivalent then it denotes $D_i = D_j$.

Researchers have proposed varied commutative cryptographic techniques [11][34][35]. Commutative cryptographic techniques could be broadly classified into symmetric type also known as private key cryptography and asymmetric type also known as public key schemes. In symmetric cryptography each custodian holds the same key for encryption and decryption. In asymmetric cryptography each custodian has independent keys for encryption and decryption. There exist a major issue with effective key distribution[31][32][33] which often results in risking the data available with the custodians. In order to address this concern the approach proposed in this paper does not consider distribution of any keys rather considers some pre decided prime integers required for key generation in commutative RSA as the only common attribute amongst the available data custodians. Further discussion of the commutative RSA algorithm its proof, cryptanalysis adopted in our approach is discussed in Appendix A of this paper.

The algorithm describing the commutative RSA agreement is given below.

Input: Prime numbers p and q pre decided amongst the data custodians.

Output: Encryption Key K_E^{Cp} and decryption key K_D^{Cp} of each data custodian involved .

Algorithm 2a:

```
1. Initialize two prime numbers $p$ and $q$ amongst all data custodian set $\mathcal{C}$
2. For each data custodian $c_{\mathcal{p}}$ in the custodian set $\mathcal{C} = \{c_1, c_2, c_3 \dots \dots . c_p\}$
3.     Calculate $n = p \times q$ and $\emptyset = (p-1)(q-1)$
4.     Compute $e$ using Algorithm 2b
5.     Compute $d = e^{-1} Mod\ \phi$
6.     Encryption Key is $K_E^{Cp} = (n\,, e_{Cp})$
```

```
7.    Decryption Key is $K_D^{Cp} = (n\,, d_{Cp})$
8. End For
```

From the above mentioned algorithm it is possible to obtain the Encryption K_E^{Cp} and Decryption Keys K_D^{Cp} at each available with data custodian to provide privacy. It is essential that no two keys available with the data custodians must be identical to provide for computational indistinguishability. This is to provide security and privacy for further secure rule transfer amongst the data custodians. To provide for computational indistinguishability the following two conditions are essential and are defined as given below

$K_E^1 \neq K_E^2 \neq K_E^3 \neq \cdots . \neq K_E^{Cp} \; \forall \, c_p \in \mathcal{C}$ and

$$K_D^1 \neq K_D^2 \neq K_D^3 \neq \cdots . \neq K_D^{Cp} \forall \, c_p \in \mathcal{C}$$

The realization of the above mentioned equations is critical and it could be observed that the parameters $n \; and \; \phi$ would be identical for all the data custodians$c_p \in \mathcal{C}$. The computation of the parameter e would differ $\forall \, c_p \in \mathcal{C}$ and would be computed using the algorithm 2b given below.

Input: $\emptyset \; i.e \; ((p-1) \times (q-1))$
Output: e to be used for encryption key
Algorithm 2b:

```
1. Initialize Pseudo Random Generator Function Generator represented by $f_{prg}(\,) = I_{prg}$ where $I_{prg}$ represents the random integer
2. Do
3.    Generate Random Integer using $f_{prg}(\,) = I_{prg}$
4.    Compute $f_{GCD}(\phi, I_{prg})$
5. While $f_{GCD}(\phi, I_{prg}) = 1$
6. $e = I_{prg}$
```

From the algorithm 2b it is clear that the random function generator is a critical function to be incorporated to provide for privacy. Computationally indistinguishable data transfers is the aim to be achieved using commutative RSA to provide for privacy preserving. For example let's consider 3 data custodians represented by $\mathcal{X}\,, \mathcal{Y} \; and \; \mathcal{Z}$ where $(\,E_x\,, D_x)\,, (\,E_y\,, D_y)$ and $(\,E_z\,, D_z)$ represent their encryption and decryption key pairs respectively. Let $\mathcal{D}$ represent the data available with $\mathcal{X} \; and \; \mathcal{Y}$. The data $\mathcal{D}$ available with $\mathcal{X} \; and \; \mathcal{Y}$ is to be provided to $\mathcal{Z}$ .For privacy provisioning custodians $\mathcal{X} \; and \; \mathcal{Y}$ send the data to $\mathcal{Z}$ by encrypting the data with their respective encryption keys represented by $E_x(D) \; and \; E_y(D)$. Privacy is provided if $E_x(D) \; and \; E_y(D)$ received by $\mathcal{Z}$ are computationally indistinguishable. i.e. $E_x(D) \neq E_y(D)$.

Phase 3: Privacy Preserved rule sharing amongst Data Custodians

Each data custodian $c_p \in \mathcal{C}$ from the two phases described above possesses their locally generated rules R_{rp} , encryption keys K_E^{Cp} and decryption keys K_D^{Cp}. In this phase we shall now discuss secure rule exchange amongst all the data custodians of the data custodian set $\mathcal{C}$ .Each custodian sends only encrypted rules to the other custodians who in turn encrypt the encrypted rules received by them using their encryption keys. This cyclic procedure is repeated in an manner such that each locally generated rules is encrypted exactly $\mathscr{p}$ times where $\mathscr{p}$ represents the total number of data custodians involved. Let $C_E R_{rp}$ represent the commutatively encrypted rule set R_{rp} of data custodian c_p , it could be defined as.

$C_E R_{rp} = \sum_{x=0}^{x=p} K_E^{Cx}\,(\,R_{rp})$

This phase is realized using Algorithm 3 given below. Let us consider a subset of the data custodian set $\mathcal{C}$ represented as $\mathcal{C}_q$ defined through the following equations.

$\mathcal{C}_q \subset \mathcal{C}$

$$\mathcal{C} = \{c_1, c_2, c_3 \dots \dots . c_p\} = \mathcal{C}_q \cup c_p$$

From the above two equations it is clear that

$\mathcal{C}_q = \{c_1, c_2, c_3 \dots \dots . c_q\}$ where $c_p \notin \mathcal{C}_q$

Input: Locally Generated C5.0 classification rules R_{rp} and Encryption Key $K_E^{Cp} = (n\,, e)$.
Output: $C_E R_{rp} \; i.e \; \mathscr{p}$ times encrypted secure rules
Algorithm 3:

```
1. For each data custodian $c_{\mathscr{p}}$ in the custodian set $\mathcal{C} = \{c_1, c_2, c_3 \dots \dots . c_p\}$
2.    Compute $K_E^{Cp}\,(\,R_{rp}) = \,R_{rp}{}^{e_{Cp}}\,(mod\; n)$
3. End For
4. For each locally encrypted dataset $K_E^{Cp}\,(\,R_{rp})$ in the custodian set $\mathcal{C}_q = \{c_1, c_2, c_3 \dots \dots . c_q\}$ where $c_p \notin \mathcal{C}_q$
5.    Custodian $Cp$ sends $K_E^{Cp}\,(\,R_{rp})$ to each custodian $\mathcal{C}_q$
6.    Compute $K_E^{Cq}\left(K_E^{Cp}\,(\,R_{rp})\right) = K_E^{Cp}\,(\,R_{rp})^{e_{Cq}}\,(mod\; n)$
7. End For
```

On applying algorithm 3 each of the data custodians $c_p \in \mathcal{C}$ would have its locally generated rules R_{rp} encrypted p times using commutative RSA encryption technique represented by. The communication cost of this phase could be represented as $O(2(p-1)\,D)$ where where O represents the communication functions and B represents the total bits transacted.

Phase 4: Construction of Combined Secure Rule Set.

This phase discusses the construction of the secure combined rule set. Let $C_E R$ represent the secure rule set defined by

$$C_E R = \left\{C_E R_{r1} \bigcup C_E R_{r2} \bigcup C_E R_{r3} \dots \dots \dots \bigcup C_E R_{rp}\right\}$$

$$C_E R = \{\sum_{x=0}^{x=p} K_E^{Cx}\,(\,R_{r1}) \bigcup \sum_{x=0}^{x=p} K_E^{Cx}\,(\,R_{r2}) \bigcup \sum_{x=0}^{x=p} K_E^{Cx}\,(\,R_{r3}) \dots \dots \dots \bigcup \sum_{x=0}^{x=p} K_E^{Cx}\,(\,R_{rp})\,\}$$

Where p represents the total number of data custodians involved in the semi-honest model considered. The secure rule set is a collection of all the rules available with each data custodian encrypted using Algorithm 3.The secure combined rule set is available with each data custodian $c_p \subset \mathcal{C}$ in the semi honest model. The algorithm for the construction of the Secure Combined Rule Set is given by Algorithm 4.

Input: $C_E R_{rp}$ commutatively rule set of data custodian c_p
Output: Secure Rule set $C_E R$

Algorithm 4 :

```
1. For each data custodian $c_p$ in the custodian set $\mathcal{C} = \{c_1, c_2, c_3 \dots \dots . c_p\}$
2.     $C_E R = C_E R \cup C_E R_{rp}$
3. End For
```

Phase 5: Construction of Combined Rule Set using commutative RSA decryption.

The penultimate phase of the model proposed in this paper discusses the use of commutative RSA decryption process at each data custodian to obtain the combined rule set represented by CR .Let c_p represent a data custodian such that $c_p \in \mathcal{C}$ who would like to construct the combined rule set. Let $\mathcal{C}_q \subset \mathcal{C}$ defined by $\mathcal{C}_q = \{c_1, c_2, c_3 \dots \dots . c_q\}$ where $c_p \notin \mathcal{C}_q$.

Then CR could be defined as

$$CR = K_D^{Cp}\left(\sum_{x=0}^{x=q} K_D^{Cx}\,(C_E R)\right)$$

$$= K_D^{Cp}\left(\sum_{x=0}^{x=q} K_D^{Cx}\left(\begin{array}{c}\{C_E R_{r1} \bigcup C_E R_{r2} \\ \bigcup C_E R_{r3} \dots \dots \dots \bigcup C_E R_{rp}\,\}\end{array}\right)\right)$$

Where $C_E R$ represents the Secure Rule Set and K_D^{Cx} represents the Decryption Key of Data Custodian x and K_D^{Cp} is the decryption key of data custodian c_p. In this phase the Combined Secure rule set $C_E R$ is decrypted by all the data custodians participating in the semi honest trust model considered. Algorithm 5 described below explains the steps involved in obtaining the Combined Rule Set CR from the combined secure rule set $C_E R$.

Input: Combined Secure Rule Set $C_E R$ and Commutative RSA Decryption Keys K_D^{Cp} of the data custodian.

Output: Combined Rule Set CR

Algorithm 5:

```
1. For each data custodian $c_q$ in the custodian set $\mathcal{C}_q = \{c_1, c_2, c_3 \dots \dots . c_q\}$
2.     For each element $C_E R_{rp}$ of the combined secure rule set $C_E R = \{C_E R_{r1} \cup C_E R_{r2} \cup C_E R_{r3} \dots \dots \cup C_E R_{rq} \cup C_E R_{rp}\}$
3.         Compute $CR^{\prime} = K_D^{Cq}(C_E R_{rp}) = (C_E R_{rp})^{d_{cq}}\ (mod\ n)$
4.         $CR^{\prime\iota} = CR^{\prime\iota} \cup CR^{\prime}$
5.     End For
6. End For
7. $CR = K_D^{Cp}(CR^{\prime\iota}) = (CR^{\prime\iota})^{d_{cp}}\ (mod\ n)$
```

From Algorithm 5 it is clear that on commutative decryption by all data custodians of the combined secure rule set $C_E R$ transforms into the combined rule set CR .The data custodian initiating the process decrypts the Combined Secure Rule Set at the end in order to preserve the privacy of the combined rule set CR. The communication cost function for this phase can be represented as $O(pB)$ where O represents the communication functions and B represents the total bits transacted.

Phase 6: Using the Combined rule set to mine the test data using C5.0

The data custodians on obtaining the combined rule set could mine their test data using the combined rule set CR and the data mining engine $\mathcal{DE}$. The data mining engine $\mathcal{DE}_p$ available with the data custodian $c_p \in \mathcal{C}$ embodies multiple functions for rule generation, classification, analysis etc. The classification function embodied within $\mathcal{DE}$ is defined as $f_c(D\,,R)$ Where D represents the data to be classified using the C5.0 algorithm and R represents the rules utilized for classification.

The combined rule set CR cannot be utilized in the classification function $f_c(D\,,R)$ as the rules obtained through the varied data custodians of the custodian set $\mathcal{C}$ need to be merged. Let R_f represent the combined final rules which primarily contain the total number of rules N followed by the rules R and could be defined as

$R_f = f_{merge}(CR)$

$R_f = f_{merge}(\{R_{r1}, R_{r2}, \dots \dots R_{rp}\})$

Where $R_{rp} = \{\, r_{1p}\,, r_{2p}, r_{3p}, \dots \dots r_{np}\,\}$ and n represents T

he total number of rules of the p^{th} data custodian

In the privacy preserving data mining model proposed in this paper $D = dt_{c\,p}$ the test data available at p^{th} data custodian and $R = R_f$ the combined final rule set obtained from Algorithm 6 discussed below.

Input: Combined Rule Set $CR, dt_{c\,p}$ test data available at p^{th} data custodian

Output: Mining result of the test data represented by $M_r^{dt_{c\,p}}$

Algorithm 6:

```
1. For each item in combined rule set $CR = \{R_{r1}, R_{r2}, \dots \dots R_{rp}\}$
2.     For Each Rule in rule set $R_{rp} = \{\, r_{1p}\,, r_{2p}, r_{3p}, \dots \dots r_{np}\,\}$
3.         $N = N + 1$
4.         $R^{\iota} = R^{\iota} \cup r_{np}$
5.     End For
6. End For
7. $R = N \cup R^{\iota}$
8. Compute $M_r^{dt_{c\,p}} = f_c(dt_{c\,p}\ , R)$ to obtain mining results
```

From the above discussed phase approach presented in this paper it is evident that privacy is preserved as no actual data is transferred amongst the data custodians only the local classification rules that are encrypted using commutative RSA are transferred providing privacy. In the next section we shall analyze the computation and the communication costs involved in the proposed model.

4. COMMUNICATION AND COMPUTATION COST ANALYSIS

This section of the paper discusses the computation and communication costs involved using the proposed privacy preserving data mining system. The computation and communication costs involved depend on the number of Data Custodians, the transactions, attributes and type of attributes available at each data custodian, the cryptographic approach adopted and the data mining algorithm incorporated into the system. The communication cost could be obtained by the number of data transfers involved and the computation costs could be obtained based on the number of encryptions, decryptions and cryptographic initializations involved. Researchers have proposed varied commutative cryptographic techniques [41][42][43].

Let $\mathcal{C}$ represent a set of data custodians containing p data custodians and $d_{c\,x}$ represents the data available with the x^{th} data custodian such that $x \in \mathcal{C}$ and $0 > x \leq p$. In this section we would compare the computational cost of Commutative RSA Proposed against Secure Lock and Asynchronous Control Polynomial. Secure Lock and Asynchronous Control Polynomial [36] consider a central server for computation and group formation. The approach discussed in this paper does not consider such central server for key distribution and computation. In addition the secure lock and Asynchronous Control Polynomial also include membership verification of the data custodians into the group.

The secure lock and ACP secure group communication scheme consider a central server for key distribution and both the schemes achieve secure group communication using cryptographic approaches. Secure Lock is assumed to utilize 1024 bit asymmetric RSA Public cryptographic scheme and ACP utilizes the advantage of an 80 bit symmetric key cryptographic scheme. Our proposed scheme utilizes a 1024 bit commutative RSA scheme for comparison efficiency [44].The computation costs[45] involved in Secure Lock, Asynchronous Control Polynomial and our proposed protocol could be summarized in the Table 1 provided below where O represents the computation function, p represents the number of data custodians, m represents the data custodians $ID's$ in the Hierarchical Access Control hierarchy [35] and K_{sl}, K_{acp}, K_{pa} represent the cryptographic keys for the Secure Lock Scheme , Asynchronous Control Polynomial
Scheme and our proposed scheme. The computation costs for encryption/decryption shown in Table 1 represent the computation involved for a single encryption/decryption computation.[45]

Table 1. Computational Cost Comparisons

Phase of Algorithm	Secure lock	Asynchronous control polynomial	Proposed algorithm
Initialization	$O(p^2 K_{sl}^2) + O(pK_{sl}^3)$	$O(p^2 K_{acp}^2)$	$O(2p) + O(2pK_{pa}^2)$
Key Computation and Derivation	$O(pK_{sl}^2) + O(K_{sl}^3)$	$O(pK_{acp}^2)$	$O(pK_{pa}^2)$
Group Membership Verification	$O(pK_{sl})$	$O(pK_{acp})$	Not Applicable
Encryption/ Decryption	$O(pK_{sl})$	$O(pK_{acp})$	$O(pK_{pa})$
Server End storage	$O(p+m)$	$O(p+m)$	Not Applicable

To prove the computational efficiency of the proposed algorithm the Secure Lock, Asynchronous Control Polynomial and our proposed scheme was developed on the Visual Studio 2010 platform using C#.Net as the programming language. The implementations were tested on an Intel Core 2 Duo 2.00 GHz CPU having 3GB of RAM. The Initialization, Key Computation and Derivation and the Group Membership Verification phases of the 3 algorithms have been considered. The results obtained are represented graphically in Figure 1 shown below. The Encryption/Decryption phase and the Server End Storage phase have been neglected for the analysis presented here. From Figure 1 it is clear that the Asynchronous Control Polynomial and the secure group communication algorithm proposed in this paper are computationally less expensive than the Secure Lock Scheme of secure group communication. The secure group communication protocol proposed in this paper provides more security and is less vulnerable to data loss as it does not consider any key exchange amongst the data custodians. Also the scheme described in this paper does not consider an central trusted server for group establishment and also for key distribution. In case of any malicious data custodians the data exchanged using Secure Lock and Asynchronous Control Polynomial is more vulnerable as the data transacted during communication is encrypted only once but in our proposed scheme the probability of the data transacted is encrypted multiple times providing for more secure means of communication.

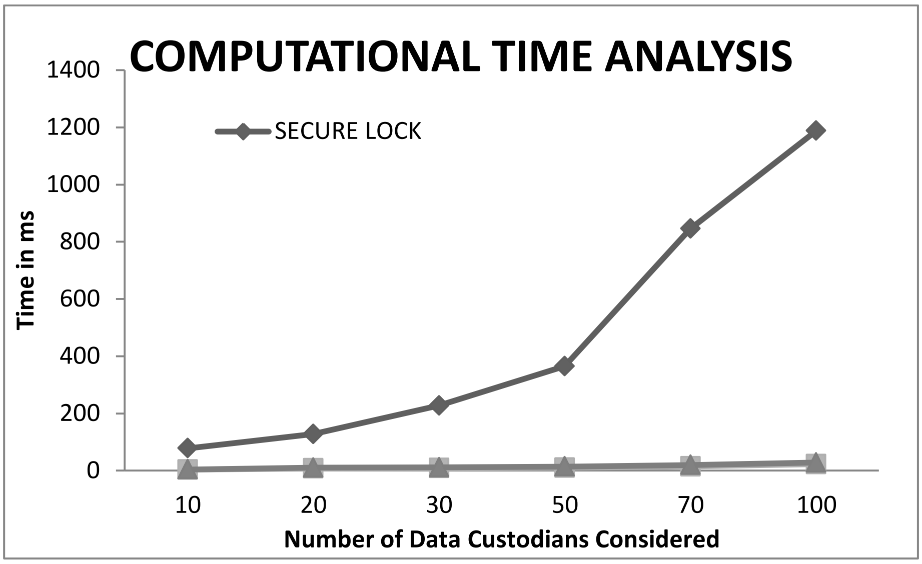

Figure 1. Computational Time Analysis

5. COMPUTATIONALLY INDISTINGUISHABLE ANALYSIS AND COMMUTATIVE NATURE PROOF OF PROPOSED MODEL

Computationally indistinguishable data transmission amongst the p data custodians considered is an ideal property of secure group communication. To prove that the transmissions amongst the data custodians are computationally indistinguishable we developed the proposed protocol for $p=2$ data custodians using C#.Net on the Visual Studio 2010 Development Platform. Fortran Libraries were utilized to handle higher mathematical computations. The data mining rules (using C5.0 algorithm) generated locally as per Algorithm 1 described in this paper for data custodian P where $P \in p$ was monitored and the data distribution graphs obtained on encryption and decryption are as shown in the figures provided below. The C5.0 algorithm used was run on the Linux Platform and the Hyperthyroid data set was considered to generate the classification rules. As p =2 there exists a maximum of p encryptions and p decryptions for the rules generated by Data Custodian P. The encryption and decryptions were carried out using commutative RSA Algorithm utilized in our approach.

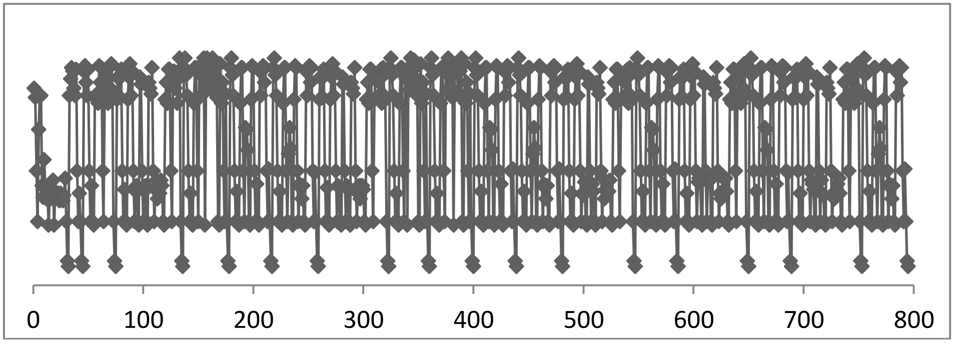

Figure 2. Data Distribution of Locally Generated Data Mining Rules for Data Custodian P

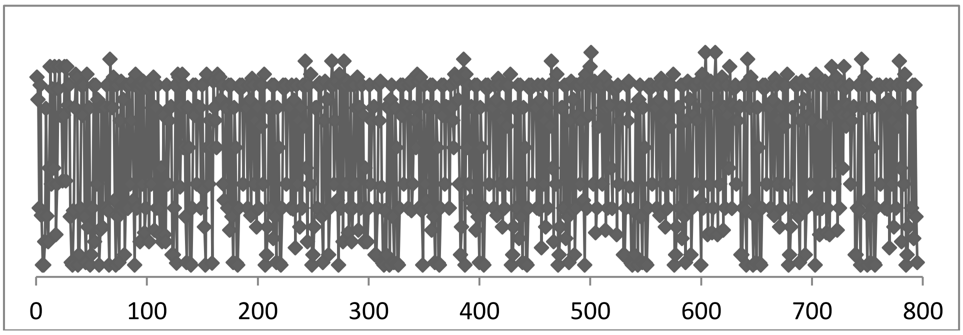

Figure 3. Data Distribution of Data Custodian P's rules Encrypted n times using Commutative RSA $(n = 1)$

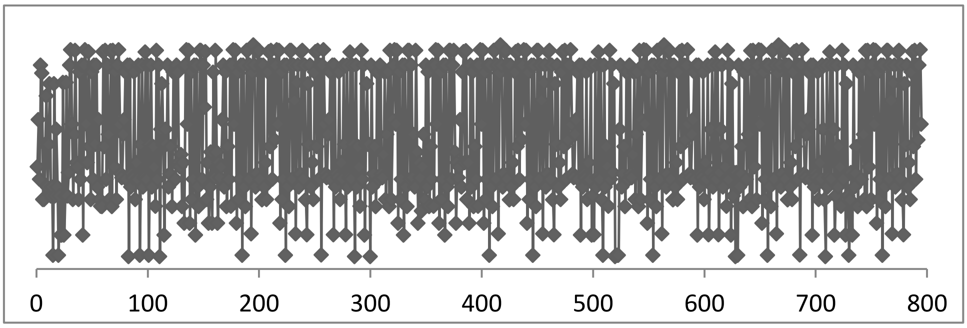

Figure 4. Data Distribution of Data Custodian P's rules Encrypted n times using Commutative RSA $(n = 2)$

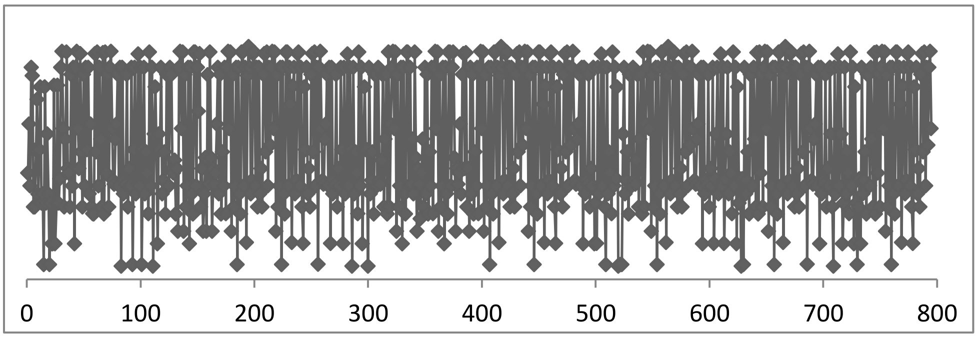

Figure 5. . Data Distribution of Data Custodian P's rules Decrypted n times using Commutative RSA $(n = 1)$

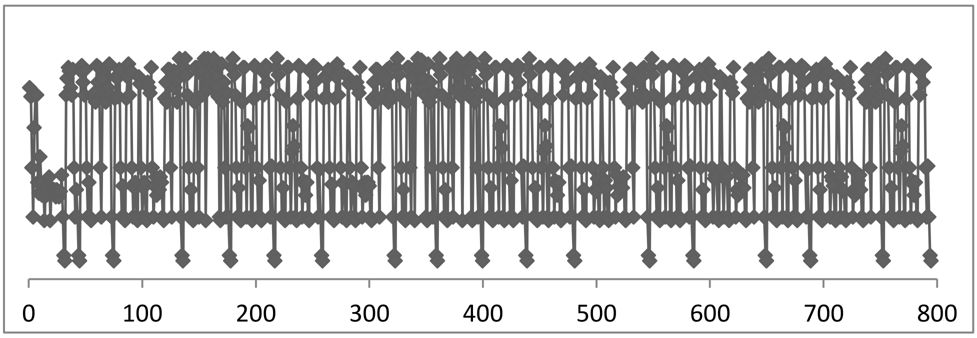

Figure 6.Data Distribution of Data Custodian P's rules Decrypted n times using Commutative RSA $(n = 2)$

Based on the above Figures it is clear that the proposed algorithm is computationally indistinguishable and it could be observed that Figure 2 is identical to Figure 6and Figure 4 is identical to Figure 5 which proves the commutative nature of the proposed system.

.

6. CONCLUSION AND FUTURE WORK

This paper introduces a new PPDM system providing prominence to the privacy and security of the horizontally portioned data available with the data custodians. Secure Multiparty Computation approach has been adopted providing privacy of the data using Commutative RSA algorithm. The system utilizes the advantages of the C5.0 algorithm for data mining, classification rule generation and for tree construction. The proposed model is described in 6 phases assuming all the data custodians abide by the semi honest trust model. The drawbacks involving key exchanges for cryptography used to secure the data transacted has been discussed and to overcome this drawback the proposed system does not consider any key exchange amongst the various data custodians involved by utilizing the advantages of commutative RSA initialization algorithm. The proposed system is compared with the existing Secure Lock and Asynchronous Control Polynomial secure group communication models in terms of the computational costs involved. The results obtained prove the efficiency of the proposed scheme eliminating the need for key exchange or key distribution discussed in this paper. The privacy and security of the data exchanged between the varied data custodians is also proved to be computationally indistinguishable through the data distribution graphs provided in this research paper.

This research represented here describes a novel PPDM model. In this paper the authors have highlighted the much desired Privacy Preserving feature and proved its efficiency. The future of the work presented here would concentrate on proving the efficiency of the C5.0 data mining model adopted for mining and achieving desired mining results.

7. REFERENCES

[1] H. Friedman and J.L. Bentley, “An Algorithm for Finding Best Matches in Logarithmic Expected Time,” ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.

[2] N.R. Adam and J.C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparison Study,” ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[3] F Matthews, Gregory J., Harel, Ofer," Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, Statistics Surveys, 5, (2011),pp 1-29 . DOI: 10.1214/11-SS074.

[4] D.G. Marks, “Inference in MLS Database,” IEEE Trans. Knowledge and Data Eng., vol. 8, no. 1, pp. 46-55 Feb. 1996.

[5] D.E. O’Leary, “Knowledge Discovery as a Threat to Database Security,” Proc. First Int’l Conf. Knowledge Discovery and Databases, pp. 107-516, 1991.

[6] M.-S. Chen, J. Han, and P. S. Yu, “Data mining: An overview from database perspective,” IEEE Trans. Knowl. Data Eng., 1996.

[7] C. Clifton and D. Marks, “Security and Privacy Implications of Data Mining,” Proc. 1996 ACM Workshop Data Mining and Knowledge Discovery, 1996.

[8] C. Clifton et al., “Tools for Privacy Preserving Distributed Data Mining,” SIGKDD Explorations, vol. 4, no. 2, 2003, pp. 28-34.

[9] Nan Zhang and Wei Zhao,"Privacy-Preserving Data Mining Systems" IEEE Computer, Vol. 40, No. 4, Pp 5258, April 2007.

[10] V.S. Verykios et al., “State-of-the-Art in Privacy Preserving Data Mining,” SIGMOD Record, vol. 33, no. 1, 2004, pp. 50-57.2001

[11] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in Proc. Int’l Cryptology Conference (CRYPTO), 2000.

[12] R. Agrawal and R. Srikant, “Privacy preserving data mining,” in Proc. ACM SIGMOD Int’l Conf. on Management of Data, 2000.

[13] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang,"Enabling Multi-level Trust in Privacy Preserving Data Mining" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2011. DOI 10.1109/TKDE.2011.124.

[14] O. Goldreich, “Secure multi-party computation,” Final (incomplete) draft, version 1.4, 2002.

[15] A.W.-C. Fu, R. C.-W.Wong, and K.Wang, “Privacy-preserving frequent pattern mining across private databases,” in Procedings of International Conference on Data Mining, 2005.

[16] Ming-Jun Xiao, Kai Han, Liu-Sheng Huang and Jing-Yuan Li,"Privacy Preserving C4.5 Algorithm Over Horizontally Partitioned Data,"Proceedings of the Fifth IEEE International Conference on Grid and Cooperative Computing, 2006

[17] J. Vaidya and C. W. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” in Proc. ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, 2002.

[18] Yanguang Shen, Hui Shao and Li Yang, "Privacy Preserving C4.5 Algorithm over Vertically Distributed Datasets",2009 IEEE,International Conference on Networks Security, Wireless Communications and Trusted Computing,DOI 10.1109/NSWCTC.2009.253, pp 446-448. 2009.

[19] J. Vaidya and C. Clifton, “Privacy-preserving k-means clustering over vertically partitioned data,” in Proc. ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, 2003.

[20] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in Proceeding of the 20th ACM Symposium on Principles of Database Systems, Santa Barbara, California.

[21] K. Chen and L. Liu, “Privacy preserving data classification with rotation perturbation,” in Proc. Int’l Conf. on Data Mining, 2005.

[22] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, “Time series compressibility and privacy,” in Proc. Int’l Conf. on Very Large Data Bases, 2007.

[23] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[24] W. Du and Z. Zhan, “Using randomized response techniques for privacy-preserving data mining,” in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[25] R. Agrawal, R. Srikant, and D. Thomas, “Privacy preserving OLAP,” in Proc. ACM SIGMOD Int’l Conf. on Management of Data, 2005.

[26] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems (IJUFKS), vol. 10, 2002.

[27] Tamir Tassa,"Secure Mining of Association Rules in Horizontally Distributed Databases".2011

[28] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), vol. 10, 2002.

[29] C. C. Aggarwal and P. S. Yu, “A condensation approach to privacy preserving data mining,” in Proc. Int’l Conf. on Extending Database Technology (EDBT), 2004.

[30] Slava Kisilevich, Lior Rokach, Yuval Elovici and Bracha Shapira,"Efficient Multidimensional Suppression for K-Anonymity",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 3, MARCH 2010

[31] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. Advances in Cryptology-EUROCRYPT 2006, pages 486–503, 2006.

[32] M. Islam , L. Brankovic. Noise Addition for Protecting Privacy in Data Mining. Proceedings of The 6th Engineering Mathematics and Applications Conference (EMAC2003), Sydney, pages 85–90. Citeseer, 2003

[33] B. Pinkas," Cryptographic techniques for privacy-preserving data mining". ACM SIGKDD Explorations Newsletter, 4(2):19, 2002.

[34] G. H. Chiou and W. Chen, “Secure broadcasting using the Secure Lock,” IEEE Transactions on Software Engineering, vol. 15, no. 8, pp. 929–934, Aug. 1989.

[35] X. Zou, Y.-S. Dai, and E. Bertino, “A practical and flexible key management mechanism for trusted collaborative computing,” Proceedings of INFOCOM’08, Phoenix, AZ, USA, pp. 1211–1219, Apr. 2008

[36] Xukai Zou, Mingrui Qi, and Yan Sui."A New Scheme for Anonymous Secure Group Communication",System Sciences (HICSS), 2011 44th Hawaii International Conference.4-7 Jan. 2011.

[37] Lakshminath R. Dondeti , Sarit Mukherjee , Ashok Samal,"Survey and Comparison of Secure Group Communication Protocols ".CiteSeerX 1999 doi=10.1.1.25.7963.

[38] Christian Cachin and Jan Camenisch,IBM Research–Zurich,"Encrypting Keys Securely",IEEE COMPUTER AND RELIABILITY SOCIETIES.1540-7993/10/.JULY/AUGUST 2010.

[39] Zhibin Zhou and Dijiang Huang,"An Optimal Key Distribution Scheme for Secure Multicast Group Communication",IEEE INFOCOM 2010.

[40] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu and Philip S. Yu, et al."Top 10 algorithms in data mining",Springer,Knowledge and Information Systems Volume 14, Number 1 (2008), 1-37, DOI: 10.1007/s10115-007-0114-2

[41] Douglas Stinson,"Cryptography: Theory and Practice",CRC Press,ISBN: 0849385210

[42] Oded Goldreich,"Foundations of Cryptography Volume I .Basic Tools",Cambridge University Press, 2004.ISBN 978-0-511-54689-1 OCeISBN.ISBN 0-521-79172-3 hardback

[43] Oded Goldreich,"Foundations of Cryptography Volume II .Basic Applications",Cambridge University Press, 2004.ISBN 978-0-521-11991-7 paperback,ISBN 978-0-521-83084-3 hardback

[44] A. K. Lenstra, “Key length,” Handbook of Information Security, Editorin-Chief, Hossein Bidgoli, vol. 2, pp. 617–635, 2005

[45] E. Bach and J. Shallit, “Algorithmic number theory, volume I: Efficient algorithms,” The MIT Press, 1996.