A Novel PPDM Protocol for Distributed Peer to Peer Information Sources

Article · January 2013

4 authors:

Kumaraswamy S
KNS Institute Of Technology
9 PUBLICATIONS 9 CITATIONS

SEE PROFILE

SH Manjula
UVCE, Bangalore University
94 PUBLICATIONS 202 CITATIONS

SEE PROFILE

Venugopal K R
University Visvesvaraya College of Engineering
925 PUBLICATIONS 3,719 CITATIONS

SEE PROFILE

Lalit M Patnaik
Indian Institute of Science
838 PUBLICATIONS 8,689 CITATIONS

SEE PROFILE

INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 4, Issue 4, July-August (2013), pp. 358-380
© IAEME: www.iaeme.com/ijcet.asp
Journal Impact Factor (2013): 6.1302 (Calculated by GISI)
www.jifactor.com

IJCET

©IAEME

A NOVEL PPDM PROTOCOL FOR DISTRIBUTED PEER TO PEER INFORMATION SOURCES

S Kumaraswamy[1], Manjula S H[1], K R Venugopal[1], L M Patnaik[2]

[1,2]Department of Computer Science and Engineering
[1]University Visvesvaraya College of Engineering, Bangalore University,
Bangalore 560 001
[2]Honorary Professor, Indian Institute of Science, Bangalore.

ABSTRACT

Cryptographic approaches are traditional and preferred methodologies used to preserve the privacy of data released for analysis. *Privacy Preserving Data Mining (PPDM)* is a new trend to derive knowledge when the data is available with multiple parties involved. The PPDM deployments that currently exist involve cryptographic key exchange and key computation achieved through a trusted server or a third party. The key computation over heads, key compromise in presence of dishonest parties and shared data integrity are the key challenges that exist. This research work discusses the provisioning of data privacy using commutative RSA algorithms eliminating the overheads of secure key distribution, storage and key update mechanisms generally used to secure the data to be used for analysis. Decision Tree algorithms are used for analysis of the data provided by the various parties involved. We have considered the C5.0 data mining algorithm for analysis due to its efficiency over the currently prevalent algorithms like C4.5 and ID3. In this paper the major emphasis is to provide a platform for secure communication, preserving privacy of the vertically partitioned data available with the parties involved in the semi-honest trust model. The proposed *Key Distribution-Less Privacy Preserving Data Mining* ($KDLPPDM$) model is compared with other protocols like Secure Lock and Access Control Polynomial to prove its efficiency in terms of the computational overheads observed in preserving privacy. The experiential evaluations proves the $KDLPPDM$ reduces the computational overheads by about 95.96% when compared to the Secure Lock model and is similar to the computational overheads observed for the Access Control Polynomial model.

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

1. INTRODUCTION

Varied parties like research organizations, government agencies, and business houses maintain data for analysis. The knowledge derived from their local data repositories is insufficient to meet their projected outcomes. Hence there exists a need for sharing data for effective data mining and better analysis. Privacy of the data available across varied parties released for analysis is of primary importance in a PPDM System. For example to curb terrorism there exists a need to analyze the data available with the immigration department of each country to understand the movement and track terrorists and suppress acts of terrorism. With the help of such cooperative systems in place, nations use the data to track and analyze the threats that terrorism poses to humanity and security of its citizens. Additionally such mechanisms can be utilized to monitor illegal immigrants. The data analysis is used for constructive or economy enhancing activities like tourism and business. Generally it is observed that the parties involved are not comfortable to disclose the entire information involved in spite of agreements and strategies in place to preserve privacy of the data. There are number of organizations like the United States Health Insurance Portability and Accountability in the United States of America and European Union Privacy Directive in Europe which have prescribed the norms to be followed prior to data released for analysis to preserve the privacy of the data [1][2][3][4][5]. To address these issues this paper proposed a novel PPDM mechanism namely $KDLPPDM$.The local data available with the parties involved in $KDLPPDM$ model is not released or shared for analysis. Instead the classification rules generated at the local parties is used for analysis thus preserving privacy of the data available with each party. The data available at each party is assumed to be vertically partitioned and all the parties involved in the proposed mechanism are semi honest in nature.

Decision Tree classification algorithms have been frequently used by researchers for accurate analysis. The data mining accuracy of the proposed system depends on the rules locally generated. The $KDLPPDM$ utilizes the C5.0 Data Mining Algorithm for accurate rule generation and classification. The use of cryptography is adopted in the $KDLPPDM$ to provide privacy to the data released for analysis. In the $KDLPPDM$ proposed in this paper the locally generated are secured using cryptographic techniques namely commutative RSA. The use of commutative RSA is adopted to overcome the drawbacks of key distribution, rekeying overheads and key compromise threats.

1.1. Motivation

The parties involved in the PPDM model governed by the semi honest trust model are not comfortable with the aspect of data release for analysis. To preserve the privacy the data researchers have adopted cryptographic techniques relying on the keys for encryption and decryptions. The keys are symmetric or asymmetric in nature. In order to derive the keys and distribute them amongst the parties the use of a third party or a secure server is considered [6][7][8]. The use of a sure third party server induces overheads related to key computation, key storage and key distribution is known as the protocol initialization phase. Post the initialization phase the data is secured using cryptography and is shared for analysis. The design of $KDLPPDM$draws its motivation targeted to eliminate the need of a third party server, minimise the network operation overheads (i.e. key storage and key distribution) and eliminate the need for the actual local data available with the parties to be released for analysis.

1.2. Contribution

The research work introduces the $KDLPPDM$ model. The $KDLPPDM$ considers the local rules generated using the C5.0 data mining algorithm to be released by each party for analysis and not the actual data. The privacy of the locally generated rules is preserved using the commutative RSA cryptographic algorithm. The $KDLPPDM$ minimizes the computational overheads associated with every privacy preserving model (namely key distribution, key storage and key computation using external entities). The privacy of the locally generated rules is preserved and is proved using the computationally indistinguishablity [9][10][11] property. The $KDLPPDM$ model proposed preserves privacy in the presence of dishonest or non trusted parties.

1.3. Organization

The remaining paper is organised as follows. The background and related work is discussed in the second and the third sections of the paper. Section four of the paper introduces the Vertically Partitioning of data. The commutative RSA algorithm is explained in fifth section of this paper. The proposed $KDLPPDM$model is discussed at length in the next section. The comparisons of the proposed $KDLPPDM$ with Secure Lock and Access Control Polynomial protocol in terms of the computational complexity is discussed in section seven of the paper. The eighth section proves the computationally indistinguishablity of the $KDLPPDM$thus preserving privacy. The conclusion and future of the research work is discussed in the last section of this paper.

2. RELATED WORK – PRIVACY PRESERVING DATA MINING

Research highlighting the benefits of data mining was introduced in the early 1980's and was adopted by business originations and other establishments a decade later. Provisioning of privacy of the data to be utilized for mining or PPDM architectures found its emergence in the beginning of the 20th century [12] [13]. There afterwards researchers proposed numerous PPDM models to secure data and facilitate data mining put forth the limitations that exist in the PPDM systems. Nan Zhang and Wei Zhao [14] discuss that PPDM systems are broadly classified based on the privacy level the systems provide. The classifications are namely Secure Multi Party Computation Techniques and Partial Information Hiding techniques.

Y. Lindell and B. Pinkas [15] proposed a PPDM system in which the parties are bound by the semi honest trust model [10]. The privacy of the data is maintained even in the presence of colluded users. The PPDM model utilized the ID3 classification mining algorithm. The major drawback of this system is that the model is limited to two parties and the classification accuracy of the ID3 algorithm is lower compared to the C4.5 and C5.0 data mining algorithm.

In secure multiparty computation techniques proposed by researchers is based on the varied data distribution techniques adopted amongst the parties involved. The data available with the parties is horizontally partitioned as considered by C. Clifton et al [13], R. Agrawal and R. Srikant [16], Yaping Li et al., [17], Ming-Jun Xiao et al., [18]. The major drawback of these systems is that the models described cannot be extended to data which is generally vertically partioned. The data mining algorithms adopted exhibit lesser classification accuracy when compared to the C5.0 data mining algorithm.

O. Goldreich [19], A.W.-C. Fu et al., [20], J. Vaidya and C. W. Clifton [21] considered vertically portioning of the data available with the parties involved. The proposed models provided for data integrity of the information available with the various parties involved. The major drawback of these systems is the reduced data mining accuracy of these systems.

The partial information hiding can be further classified into three categories namely data perturbation, k-anonymity and retention replacement. D. Agrawal and C. C. Aggarwal [22][23] , K. Chen and L. Liu [24] and S. Papadimitriou et al., [25] adopted the data perturbation technique to preserve the privacy of the data to be released for mining. The major drawback of these approaches is that the data perturbation techniques adopted effect the data mining results.

To overcome the drawbacks of data perturbation privacy preserving technique L. Sweeney et al., [26], C. C. Aggarwal and P. S. Yu [27], Slava Kisilevich et al.,[28] proposed the use of k-anonymity privacy preserving technique. The use of these techniques exhibit a high degree of data annomization and reduced mining accuracy.

W. Du and Z. Zhan[29], R. Agrawal, R. Srikant, and D. Thomas[30] in their papers introduced retention replacement techniques to preserve privacy. The retention replacement techniques elements with a certain probability are retained as it is or the elements are replaced based on the type of elements and probability distribution function adopted for that type of element. The major drawback of these systems is that such privacy preserving techniques can be adopted only for data that is continuous in nature. In general partial information hiding technique is not capable of effectively performing in the presence of colluded users hence the approach has not been adopted in the proposed approach.

Knowledge extraction or data mining is a vital operational requirement of any successful PPDM system. The $KDLPPDM$ protocol presented in this paper adopts C5.0 decision tree algorithm to the purpose of mining [31][32][33][34]. The use of C5.0 classification tree algorithm is preferred over its predecessors decision tree algorithms like ID3 used by J. Vaidya and C. Clifton [21], Xindong Wu et al., [32] and C4.5 adopted by Ming-Jun Xiao et al.,[18], Yanguang Shen et al., [35] due to the mining accuracy it provides[32].

In addition to the above mentioned techniques there have been efforts to provide data privacy utilizing introduction of noise to preserve privacy Po-Hsun Sung et al., [36], C. Dwork et al., [37].Ineffective de-noising techniques and inability to obtain original data is the major drawback of this approach. The data available with the parties have been secured by using specific anonomization techniques introduced by Matthews et al., [5]. The annonomization technique adopted by Matthews et al., [5] effect the mining results hence such annomizations techniques have not been adopted.

The use of cryptographic techniques to secure data transmitted for mining by Yaping Li [17] , Ming-Jun Xiao et al., [18] , R. Agrawal and R. Srikant [16] , B. Pinkas [38] have found to be very effective and do not effect the mining accuracy as desired in PPDM systems. The major drawback of these systems is the use of traditional data mining algorithms which render limited mining results.

The novelty of the proposed PPDM protocol is evident as it not only provides security using commutative algorithms [39][11] and eliminates the overheads arising due to key distribution , key storage overheads and re-keying[40][41][42].

3. BACKGROUND WORK

G. H. Chiou et al., [6] proposed the “Secure Lock” secure group communication protocol. This protocol considers an external third party server for key computation and key distributation. The secure lock protocol is based on the asymmetric cryptographic protocol and utilizes a lock based mechanism to secure the data transactions over the network. The major drawback of the secure lock protocol is that the protocol is computationally heavy and considers a third party server for the initialization phase. To overcome the drawbacks of the computationally cumbersome Secure Lock protocol Xukai Zou et al., [7] [8] proposed the Access Control Polynomial protocol. The Access Control Polynomial introduces a novel key management scheme adapted to support multi party group communications. The Access Control Polynomial Protocol is computationally lighter than the Secure Lock Mechanism as it adopts the symmetric cryptographic techniques to preserve privacy and data integrity. The Access Control Polynomial considers an external central server for the protocol initialization which contributes to key computation, key storage, key management and key exchange overheads there by reducing its efficiency.

J. Vaidya et al., [43][21] and R. Agrawal et al., [16] considered the ID3 decision tree data mining algorithm for analysis of the data released by the parties in the PPDM model. The drawback ID3 data mining algorithm is that it cannot handle missing parameters in datasets, continuous data, provides no support for pruning. To overcome the drawbacks researchers Ming-Jun Xiao et al., [18] and J. Vaidya et al., [21][43] utilized the C4.5 algorithms to attain higher classification accuracy and handle continuous data. The C4.5 algorithm was improved by J. Ross Quinlan [31][32]33][34] to obtain higher classification accuracy and reduce the execution time.

4. VERTICALLY PARTITIONED DATA

This paper highlights a PPDM protocol when the data available with the parties involved is vertically partitioned. Vertically partitioned data or heterogeneous distribution of data occurs when the parties involved possess data of independent attributes and all the attributes available with the parties composite the complete transactions to be utilized for mining. Let us consider a set of p parties involved in a semi honest trust model represented by the set $P_p = \{ p_{p1} , p_{p2} , \ldots p_{pp} \}$
Let the data to be utilized for Mining M_d can be represented as

$M_d = \{ Md_{t1} , Md_{t2}, \ldots .. Md_{tt} \}$

Where t Represents the total number of transactions and
Md_{tt} Represents the each transaction of the mining data M_d
Each Transaction Md_{tt} is represented as

$$Md_{tt} = \{ Att_1, Att_2, \ldots . Att_a\}$$

Let the data available with the p party be represented as

$$p_p Md = \{Att_1, Att_2, \ldots . Att_p\}$$

Where $Att_1 , Att_2 \ldots . Att_p$ Represents the attributes available with Party p_p
The M_d Data is said to be vertically partitioned if

$M_d = \{p_1 Md \cup p_2 Md \cup p_3 Md \cup \ldots \ldots p_p Md \}$

For example let us analyze the effects of pollution on heart diseases. Data is available with the Pollution control Board and the health ministry. The pollution control board provides an area wise pollution parameters and the health departments provides the data related to the personal in that area having heart related ailments. With the help of these two data it is possible to study the effect of pollution on heart ailments.

5. COMMUTATIVE RSA

Provision of privacy of the data released for mining is an important functionality to be provided in any PPDM system. Cryptography is the preferred means to provide for privacy of the data. In this paper we have adopted the commutative property of the RSA Algorithm for provisioning of privacy. We shall now discuss the construction proof and the commutative nature of the RSA algorithm adopted in the $PPDM$. The proposed protocol assumes that the parties involved pre exchange two prime numbers p and q prior for key generation amongst themselves as a prerequisite for the construction of the PPDM system considered. The prime numbers selected are randomly decided such that $p > q$.

Let us consider 2 parties named *Aruna* and *Jahnavi* to have exchanged two prime numbers p and q such that $p > q$.The RSA is an asymmetric cryptographic algorithm so *Aruna* and *Jahnavi* generate the encryption and decryption keys based on the following equations. The encryption key is represented as (n, e) and the decryption key as (n, d).

$n = p \ \times q$

$\emptyset = \emptyset(p) \times \ \emptyset(q) = (p-1) \times (q-1)$

It's observed that the n obtained by *Aruna* and *Jahnavi* is similar, public and does not reveal the pre computed prime numbers p and q. Now to obtain e each party selects a large random number such that it is a co prime of $\emptyset$. I.e. the following equation must be satisfied to find the co prime.

$$GCD\ (e\ ,\emptyset) = \ 1$$

The parameter d is computed using the following equation.

$$d = \ e^{-1}\ Mod\ \emptyset$$

Let us consider a data L . The encryption of the data L using the RSA Algorithm is represented as $X = \ L^{e}\ Mod(n)$ And the decryption operation is represented as

$$\mathrm{Y} = \ \mathrm{X^{d}}\ \mathrm{Mod}\ (n)$$

Where Y represents the decrypted data.

Let us consider that the encryption keys of *Aruna* and *Jahnavi* are $(n, e_s), (n, e_d)$ and the decryption keys are represented as $(n, d_s), (n, d_d)$. To prove the commutative nature of the algorithm we need to prove that for the considered data represented as L, if *Aruna* encrypts the data first and then *Jahnavi* encrypts the data the resultant is equivalent to resultant obtained when *Jahnavi* encrypts the data L and then *Aruna* encrypts the data again. Let E represent the commutative RSA operation and the data considered to be represented by L.

$E_d(\ E_s) \equiv \ E_s(\ E_d)$

$E_d(\ L^{e_s}\ Mod(n)) \equiv \ E_s(\ L^{e_d}\ Mod(n)\ )$

$(\ L^{e_s\, e_d}\ Mod(n)) \equiv \ (\ L^{e_d e_s}\ Mod(n)\ )$

6. KEY DISTRIBUTION-LESS PRIVACY PRESERVING DATA MINING (KDLPPDM) SYSTEM

This section of the paper discusses the proposed $KDLPPDM$. The protocol adopts a three step approach apart from the system initialization step. In the system initialization step each party generates mining rules locally using the C5.0 algorithm on the pre classified data available with each party. The initialization step additionally discusses commutative RSA algorithm initialization wherein each party computes their encryption and decryption keys.

The List of symbols used in the paper shown in the Table1 provided below.

Table 1. Notations used in the Algorithms

SYMBOLS	DEFINITION
$\mathcal{P}ty$	A Set of parties involved in the PPDM.
pty_{n}	The n th party of the set $\mathcal{P}ty$.
$\mathcal{P}ty_{\mathrm{rem}}$	Set of parties excludingpty_{n}or initiator party.
$\mathcal{D}t$	Vertically partitioned data of n parties.
dt_{n}	Pre classified data.
$\mathrm{Rl}_{\mathrm{dt}}$	Classification rules obtained from the data dt.
Cl_{r}	Rules Classification set.
Rl_{n}	Rule Set.
$\mathcal{R}_{\mathrm{n}}^{\mathrm{C5.0}}$	Classification rules obtained from C5.0 algorithm.
$\mathcal{R}_{\mathrm{pool}}$	Combined classification rules.
$\mathcal{T}_{\mathrm{Init}}$	Computation overhead in PPDM protocol initialization.
$\mathcal{T}_{\mathrm{i}}$	Computation overhead in computing n,phi,GCD.
$\mathcal{T}_{\mathrm{kc}}$	Computation overhead in computing Encryption and Decryption keys.
pty_{int}	Party initiator used to construct combined rule pool.
$\mathcal{SR}_{\mathrm{pool}}$	Secure rule pool.
$\mathrm{rl}_{\mathrm{1mptyn}}$	Maximum number of rules obtained from party pty_{n}
$\mathrm{O(n)}$	Represent the communication function
$\mathcal{C}\mathrm{l\_DM\_Rlst}_{\mathrm{n}}^{\mathrm{C5.0}}$	Classification algorithm function of the C5.0 data mining algorithm.
$\mathrm{f}_{\mathrm{rlmrg}}^{\mathrm{C5.0}}$	Combined rule file generation function.
$\mathrm{f}_{\mathrm{clss}}^{\mathrm{C5.0}}$	C5.0 Classification function.
$\mathcal{Z}$	Total data bits transacted.
ddt_{n}	Unclassified data available with $pty_{\mathrm{n}} \in \mathcal{P}ty$
$\mathcal{C}\mathrm{omm}_{\mathrm{Step2}}$	Communication cost of step 2
$\mathcal{R}_{\mathrm{n}}^{\mathrm{C5.0}}$	Classification rules for n parties
$\mathcal{R}_{\mathrm{ntmp}}^{\mathrm{C5.0}}$	Classification rules
$\mathrm{rl}_{\mathrm{1mptyn}}$	The max number of rules obtained from party pty_{n} .
$\mathcal{R}_{\mathrm{cnt}}$	Rules count.
$(n_{pty_{\mathrm{n}}} , d_{pty_{\mathrm{n}}})$	Decryption key of each party.
$n_{pty_{\mathrm{n}}} , e_{pty_{\mathrm{n}}})$	Encryption key of each party.
p , q	Two prime numbers p and q where $p > q$.
$\mathcal{D}t_{\mathrm{n}}$	Set of r number of transctions.
$tr_{\mathrm{n}r}$	r th number transaction or total number of transactions in set $\mathcal{D}t_{\mathrm{n}}$.
$at_{\mathrm{n}r\mathrm{t}}$	t^{th} unique attribute.

The steps of the $KDLPPDM$ are as follows

i. The provisioning of privacy of the locally generated rules and construction of the secure combined rule pool. The secure combined rule pool is a collection of the locally generated rules by each party in an encrypted form to preserve privacy.

ii. Obtaining the combined rule pool to use these rules for mining and analysis. In this step one party is initialized as the initiator who propagates the secure rule pool amongst the various parties. The parties involved decrypt the secure rule pool using the commutative RSA decryption key. The initiator who decrypts the secure rule set finally obtains the combined rule set.

iii. The analysis and the mining result of the unclassified data using the C5.0 classification algorithm.

We now discuss the preliminary notations used to describe the model, the initialization of the PPDM protocol and the three step approach adopted in the realization of the $KDLPPDM$ is discussed.

6.1. Preliminary Notations used in KDLPPDM

Let us consider the set $\mathcal{Pty}$ represents the n parties involved in the PPDM system considered. It is assumed that the n parties involved agree to participate in a Semi-Honest Trust Model. The set $\mathcal{Pty}$ is defined as follows

$$\mathcal{Pty} = \{pty_1, pty_2, pty_3, \dots \dots \dots\ pty_n\}$$

The set $\mathcal{Pty}$ can be defined as

$$\mathcal{Pty} = \{pty_1, pty_2, pty_3, \dots \dots \dots\ pty_m, pty_n\}$$

Where $m \neq n$ and $1 \leq m < n$

Let $\mathcal{Pty}_{rem}$ represent a set of parties in the PPDM system excluding party pty_n , defined by

$$\mathcal{Pty}_{rem} = \mathcal{Pty} \cap pty_n$$

$$\mathcal{Pty}_{rem} = \{pty_1, pty_2, pty_3, \dots \dots \dots\ pty_m, pty_n\} \cap pty_n$$

$$\mathcal{Pty}_{rem} = \{pty_1, pty_2, pty_3, \dots \dots \dots\ pty_m\}$$

Let the set $\mathcal{Dt}$ represent the vertically partitioned data available with all the n parties involved in the PPDM system defined by

$$\mathcal{Dt} = \{dt_1 \cup dt_2 \cup dt_3 \dots \dots \cup dt_n\}$$

The proposed PPDM protocol assumes that each party contains pre classified data and unclassified data. The data $\mathcal{Dt}$ consists of $\mathcal{Tr}$ transactions which contain $\mathcal{At}$ attributes. The data available with the parties is vertically partitioned .i.e. each of the n parties have $\mathcal{Tr}$ transactions set available with them but no two parties have the same attribute appearing in their dataset. The data available with the n^{th} party is defined as

$$\mathcal{Dt}_n = \{tr_{n1}, tr_{n2}, tr_{n3}, \dots .\ tr_{nr}\}$$

And $\mathcal{Dt}_n \neq \emptyset$ Where r represents the total number of transactions
Each transaction tr_{nr} contains t unique attributes represented by

$$\mathcal{Tr}_{nr} = \{at_{nr1}, at_{nr2}, \dots\ at_{nrt}\}$$

The data available with the n^{th} party represented as

$$\mathcal{Dt}_n = \{\{at_{n11}, at_{n12}, \dots\ at_{n1t}\}, \{at_{n21}, at_{n22}, \dots\ at_{n2t}\}, \dots .\{at_{nr1}, at_{nr2}, \dots\ at_{nrt}\}\}$$

Where $\mathcal{D}t_n \subset \mathcal{D}t$. As $at_{nrt} \in \mathcal{T}r_{nr}$ and $\mathcal{T}r_{nr} \in \mathcal{D}t_n$ .We can state that $at_{nrt} \in \mathcal{D}t_n$ and $at_{nrt} \notin \mathcal{D}t \cap dt_n$ as the data available with each party of the set $\mathcal{P}ty$ is vertically partitioned. The data set $\mathcal{D}t$ represents the pre classified data set dt_n available with each party $pty_n \in \mathcal{P}ty$. Each party pty_n contains a data set ddt_n which represents an unclassified data which needs to be classified. As the locally generated rules provide lesser mining accuracy [44] the parties of the set $\mathcal{P}ty$ agree to share the locally generated rules to achieve a higher degree of mining accuracy in an semi-honest trust model[13][17][38].

6.2. KDLPPDM System Initialization

The PPDM system discussed in this paper considers that the data $\mathcal{D}t$ is vertically partitioned and the pre classified data dt_n available with the $pty_n \in \mathcal{P}ty$ is used to generate the classification rules using the C5.0 data mining algorithm. The C5.0 algorithm is preferred owing to its higher classification accuracy when compared to the commonly used classification tree algorithms like ID3[21][32] and C4.5[13][12][43]. Let $f_{cl}^{C5.0}(dt, Rl_{dt})$ represent the C5.0 classification function and $dt \in \mathcal{D}t$ and Rl_{dt} represents the classification rules obtained from the data dt. The rule generation function based on the pre classified data dt and the classification set

$Cl_r = \{cl_1, cl_2, \dots cl_c\}$ is defined as $\quad f_{rlgen}^{C5.0}(dt, Cl_r) = Rl$

Where Rl the rule set constructed is based on the data dt and the classification set Cl_r.

Algorithm : *KDLPPDM* Initialization

Input: Two prime numbers p and q where $p > q$
Output: Encryption key =(n_{pty_n} , e_{pty_n}), Decryption key = (n_{pty_n} , d_{pty_n}) Local classification rules, Rl_n *for each party*

```
Require 2 prime numbers $p$ and $q$ where $p > q$
For each party $pty_n \in \mathcal{P}ty$
    Initialization of $\mathcal{P}p_{pty_n}$ Privacy Preserving Function
        $p_{pty_n} = p$
        $q_{pty_n} = q$
        Compute $n_{pty_n} = p_{pty_n} \times q_{pty_n}$
        Compute $\emptyset_{pty_n} = \emptyset(p_{pty_n}) \times \emptyset(q_{pty_n}) = (p_{pty_n} - 1) \times (q_{pty_n} - 1)$
        Obtain $e_{pty_n} \mid \mathcal{GCD}(e_{pty_n}, \emptyset_{pty_n}) = 1$
        Obtain $d_{pty_n} = e_{pty_n}^{-1} \mathcal{M}od(\emptyset_{pty_n})$
        Party $p_{pty_n}$ encryption key = ( $n_{pty_n}$ , $e_{pty_n}$ )
        Party $p_{pty_n}$ decryption key = ( $n_{pty_n}$ , $d_{pty_n}$ )
    End initialization of $\mathcal{P}p_{pty_n}$ Privacy Preserving Function
    Initialization of $f_{rlgen}^{C5.0}(dt, Cl_r)$ C5.0 Rule Generation
        Obtain Pre classified data set of $pty_n$ $dt_n$
        Obtain Classification Set $Cl_r$
        Compute local classification rules $f_{rlgen}^{C5.0}(dt_n, Cl_r) = Rl_n$
    End Rule Generation
End for each.
```

The rule set Rl generated locally are released by all the parties of the set $\mathcal{Pty}$ for analysis. Preserving the privacy of the locally generated rule set is achieved by adopting the advantages of Commutative RSA. There is a need to generate the encryption and decryption keys for each party. To obtain the keys using commutative RSA the parties exchange two prime numbers p and q where $p > q$ .Let $\mathcal{P}p_{pty_n}$ represent the privacy preserving function incorporated in the proposed PPDM system , using commutative RSA. The greatest common divisor function is denoted as $\mathcal{GCD}(x, y)$. The system initialization phase is described by the algorithm given above.

On completion of the initialization each party pty_n of the set $\mathcal{Pty}$ possesses its encryption and decryption keys for preserve the privacy of the rules Rl_n generated using the C5.0 algorithm. The locally generated rules are utilized in the construction of the secure rule pool using commutative RSA as discussed in the next section of this paper. The objective of the proposed PPDM protocol is be stated as the construction of the combined rule pool $\mathcal{R}pool$ defined by preserving privacy

$$\mathcal{R}_{pool} = \{Rl_1 \cup Rl_2 \cup Rl_3 \; \ldots . \cup Rl_n \} \; \forall \, n$$

Where n represents party $pty_n \in \mathcal{Pty}$. The $\mathcal{R}_{pool}$ is utilized to obtain the mining results on the unclassified data setddt_n.

The locally generated rules Rl_n are provided by each party for analysis. The actual data dt_n available with party n is not compromised and not shared amongst the varied parties involved even though there exists a cryptographic protocol (Commutative RSA) in place to provide privacy like in the prevalent systems [45][46] in place.

The computation overheads represented involved in the PPDM protocol initialization is defined as follows

$$\mathcal{T}_{Init} = \mathcal{T}_i + \mathcal{T}_{kc}$$

Where $\mathcal{T}$ represents the computational overhead observed in terms of the computational times involved and $\mathcal{T}_i$ represents the computational overhead applicable in computing n , $\emptyset$ and the $\mathcal{GCD}$. $\mathcal{T}_{kc}$ represents the computational overhead in computing the encryption and decryption keys. Let $O(n)$ represent the time complexity function involved in solving n bit operations. The computational overhead is defined as follows

$$\mathcal{T}_{Init} = O(2p) + O\left(2pK_{pa}^2\right) + O\left(pK_{pa}^2\right)$$

6.3. Step 1: Privacy Provisioning of the Classification Rules and Construction of the Secure Rule Pool

The foremost step of the proposed PPDM protocol is targeted towards the construction of the secure rule pool represented by $\mathcal{SR}_{pool}$ and the privacy preserving feature of the locally generated rules. To provide for privacy each party releases the locally generated rules Rl_n after encrypting using its encryption key (n_{pty_n} , e_{pty_n}). All the other parties encrypt the rules Rl_n using their respective encryption keys. The commutative encryption function for data x using the encryption key (n, e) is defined as

$$\mathcal{E}\,(x) = \; x^e \; \mathcal{M}od(n)$$

Extending the above definition to a multi part scenario the encryption function performed by party n defined as

$$\mathcal{E}_n\,(x) = \; x^{e_{pty_n}} \; \mathcal{M}od(n)$$

Let $\mathcal{SR}_{pool}$ represent the secure rule pool to be constructed. The secure rule pool defined as $\mathcal{SR}_{pool} = \{SRl_1 \cup SRl_2 \cup SRl_3 \; \ldots . \cup SRl_n \} \; \forall \, n$

Where n represents the total number of parties involved in the PPDM model and SRl_n represents the secured rule set of $\mathcal{pty}_n$ obtained after n number of encryptions.

The privacy provisioning of the locally generated rule set is achieved using the algorithm mentioned below

```
Algorithm Name : Secure Rule Pool Construction $\mathcal{SR}_{pool}$

Input: Encryption key = ( $n_{\mathcal{pty}_n}$ , $e_{\mathcal{pty}_n}$ ) ,Local classification rules $Rl_n$ of each party
Output: Secure Rule Pool $\mathcal{SR}_{pool}$

Initialize $\mathcal{SR}_{pool} = \emptyset$
For each $\mathcal{pty}_n \in \mathcal{Pty}$
    Compute $SRl_n = \mathcal{E}_n\,(Rl_n) = Rl_n^{e_{\mathcal{pty}_n}}\ \mathcal{Mod}(n)$
    Compute $\mathcal{Pty}_{rem} = \mathcal{Pty} \cap \mathcal{pty}_n = \{\mathcal{pty}_1, \mathcal{pty}_2, \mathcal{pty}_3, \dots\dots\dots\ \mathcal{pty}_m\}$ where $m \neq n$
    For Each $\mathcal{pty}_m \in \mathcal{Pty}_{rem}$
        Compute $SRl_{mn} = \mathcal{E}_{mn}\,(SRl_n) = SRl_n^{e_{\mathcal{pty}_m}}\ \mathcal{Mod}(n)$
    End For
    $\mathcal{SR}_{pool} = \mathcal{SR}_{pool} \cup SRl_{123\dots mn}$
End For each
```

From the above algorithm it is evident that the rules available with each party are encrypted n times. To construct the secure rule pool there exists an communication cost involved owing to the transfers of the secure rules amongst the parties involved. The communication cost $Comm_{Step1}$ of the first step defined as

$$Comm_{Step1} = O(2(n-1)\,\mathcal{Z})$$

Where$O(n)$ represents communication function n represents the parties involved and $\mathcal{Z}$ represents the total data in bits transacted.

This is a very critical step involved as this step provides the privacy of the data using commutative cryptography approaches. In the case of a malicious party involved the system is designed such that the even on obtaining any data, the obtained data is in an cryptic format with zero knowledge as to how many times the data is encrypted there by providing privacy. More over the data transacted amongst the parties is computationally indistinguishable. For better understanding let's consider parties denoted by $\mathcal{P}\,,\mathcal{Q}\ and\ \mathcal{R}$ and $(\,E_{\mathcal{p}}\,,D_{\mathcal{p}})\,,(\,E_{\mathcal{q}}\,,D_{\mathcal{q}})$ and $(\,E_{\mathcal{r}}\,,D_{\mathcal{r}})$ represent their encryption and decryption functions respectively. Let the data available with $\mathcal{P}\ and\ \mathcal{Q}$ be represented as $\mathcal{X}$. The data $\mathcal{X}$ available with $\mathcal{P}\ and\ \mathcal{Q}$ is to be provided to $\mathcal{R}$ .For privacy provisioning the parties $\mathcal{P}\ and\ \mathcal{Q}$ provide the data to $\mathcal{R}$ by encrypting the data with their respective encryption functions represented by $E_p(\mathcal{X})\ and\ E_q(\mathcal{X})$. Privacy is provided if $E_p(\mathcal{X})\ and\ E_q(\mathcal{X})$ received by $\mathcal{R}$ are computationally indistinguishable. i.e. $E_p(\mathcal{X}) \neq E_q(\mathcal{X})$. The computational indistinguishablity of the proposed PPDM protocol is discussed in the future sections of the paper.

6.4. Step 2: Construction of the Combined Rule Pool maintaining Privacy

This step of the proposed PPDM protocol is targeted towards the construction of the combined rule set from the secure rule set constructed using step 1

Decryption of the secure rule pool at each party is achieved using the Commutative RSA algorithm utilizing their respective decryption keys (n, d). The commutative decryption function of the encrypted data x is defined as

$$\mathcal{D}\,(x) = \; x^{d}\, \mathcal{M}od(n)$$

Extending the above definition to a multi party communication scenario for a party n defined as

$$\mathcal{D}_n\,(x) = \; x^{d_{pty_n}}\, \mathcal{M}od(n)$$

Let Rl_n represent the C5.0 classification rules of the party $pty_n \in \mathcal{P}ty$. The combined rule pool is defined as

$$\mathcal{R}_{pool} = \{Rl_1 \cup Rl_2 \cup Rl_3 \; \dots . \cup Rl_n\}$$

From the *combine rule pool* algorithm it is clear that the initiator pty_{int} propagates the secure rule pool $\mathcal{SR}_{pool}$ amongst all the parties $\mathcal{P}ty_{rem}$ bound in the considered $KDLPPDM$ model. Each party decrypts the secure rules of the secure rule pool and then sends it to the next party in such a way that the initiator decrypts the secure rule pool last. The initiator receives the secure rule pool decrypted exactly $(n-1)$ times. The final decryption of the secure rule pool performed by the initiator provides the combined rule pool $\mathcal{R}_{pool}$.

The algorithm adopted in the construction of the combine rule pool is as mentioned below.

Algorithm Name : Combined Rule Pool Construction$\mathcal{R}_{pool}$

Input: Decryption key = (n_{pty_n} , d_{pty_n})of each party and Secure Rule Pool $\mathcal{SR}_{pool}$
Output: Combined Rule Pool $\mathcal{R}_{pool}$

```
Initialize $\mathcal{R}_{pool} = \emptyset$
Initialize initiator $pty_{int} \in \mathcal{P}ty$
Compute $\mathcal{P}ty_{rem} = \mathcal{P}ty \cap pty_{int} = \{pty_1, pty_2, pty_3, \dots\dots\dots\ pty_m\}$
where $m \neq int$
For each $pty_m \in \mathcal{P}ty_{rem}$
    Initialize $Temp_m\mathcal{R}_{pool} = \emptyset$
        For each secure rule $SRl_n \in \mathcal{SR}_{pool}$
            Compute $\mathcal{D}_m\,(SRl_n) = SRl_n{}^{d_{pty_m}}\,\mathcal{M}od(n)$
            $Temp_m\mathcal{R}_{pool} = Temp_m\mathcal{R}_{pool} \cup \mathcal{D}_m\,(SRl_n)$
        End For
    $\mathcal{SR}_{pool} = Temp_m\mathcal{R}_{pool}$
End For
Initialize $Temp_{init}\mathcal{R}_{pool} = \emptyset$
For each secure rule $SRl_n \in \mathcal{SR}_{pool}$
    Compute $\mathcal{D}_{init}\,(SRl_n) = SRl_n{}^{d_{pty_{init}}}\,\mathcal{M}od(n)$
    $Temp_{init}\mathcal{R}_{pool} = Temp_{init}\mathcal{R}_{pool} \cup \mathcal{D}_{init}\,(SRl_n)$
End For
$\mathcal{R}_{pool} = Temp_{init}\mathcal{R}_{pool}$
```

The communication cost $\mathcal{C}omm_{Step2}$ of the second step is defined as

$$\mathcal{C}omm_{Step2} = O(2(n-1)\,\mathcal{Z})$$

Where$O(n)$ represents communication function n represents the parties involved and $\mathcal{Z}$ represents the total data in bits transacted. The bits transferred $\mathcal{Z} \propto \mathcal{SR}_{pool}$.

Though this step any party is capable of construction the combined rule pool $\mathcal{R}_{pool}$ securely without worrying about privacy as the transacted data is always in a cryptic form providing security even in the presence of malicious parties. Moreover all the parties involved are unaware of the encryption and decryption keys utilized in the $KDLPPDM$ model as there is no key exchange involved. The computational indistinguishablity of this step is proved in the results and discussion section of this paper.

6.5 Step 3: Analysis using the C5.0 Algorithm using the Combined Rule Pool

The first two steps of the $KDLPPDM$ model concentrated on provisioning of the privacy of the classification rules exchange. This step discusses the data mining algorithm adopted for analysis of the unclassified data available with the party. Once the rules generated by all the parties are accumulated, these rules are combined to generate stronger classification rules utilized in mining.

Let us consider ddt_n represents the unclassified data available with $\mathcal{pty}_n \in \mathcal{Pty}$. Let the combined rule file generation utilizing all the n rules of the combined $\mathcal{R}_{pool}$ be defined as

$$\mathcal{R}_n^{C5.0} = f_{rlmrg}^{C5.0}\left(\mathcal{R}_{pool}, n\right)$$

$$\mathcal{R}_n^{C5.0} = f_{rlmrg}^{C5.0}(\{Rl_1 \cup Rl_2 \cup Rl_3 \ldots . \cup Rl_n\}, n)$$

$$\mathcal{R}_n^{C5.0} = f_{rlmrg}^{C5.0}\left(\left\{\left\{rl_{11}, rl_{21} \ldots ., rl_{1mpty1}\right\} \cup \left\{rl_{12}, rl_{22} \ldots ., rl_{1mpty2}\right\} \cup \ldots . \cup \left\{rl_{1n}, rl_{2n} \ldots ., rl_{1mptyn}\right\}\right\}, n\right)$$

From the above definition it is clear that the combined rule file $\mathcal{R}_n^{C5.0}$ embodies all the rules of $\mathcal{pty}_1$ to $\mathcal{pty}_n$. The max number of rules obtained from party $\mathcal{pty}_n$ is represented as rl_{1mptyn}.Where $f_{rlmrg}^{C5.0}$ represents the combined rule file generation function.
Let the classification algorithm function of the C5.0 data mining algorithm be represented as

$$\mathcal{C}l\_DM\_Rlst_n^{C5.0} = f_{clss}^{C5.0}(\mathcal{R}_n^{C5.0}, ddt_n)$$

Where $f_{clss}^{C5.0}$ represents the C5.0 classification function. The classification function considers the classification rules $\mathcal{R}_n^{C5.0}$ and the unclassified data ddt_n available with party n as inputs. The output of the function is denoted as $\mathcal{C}l\_DM\_Rlst_n^{C5.0}$.
The classification data $\mathcal{C}l\_DM\_Rlst_n^{C5.0}$ represented in the form of decision trees is utilized for analysis of the unclassified data ddt_n.

The $KDLPPDM$ model discussed in this paper preserves the privacy of the data released for analysis and provides a rule based classification analysis platform using the advanced C5.0 decision tree algorithm [31][32][33][34]. Data available with each party involved is kept secure and no party releases that data for analysis instead only the rules generated are exchanged for analysis.

The data mining process of the $KDLPPDM$ model is achieved on the basis of the algorithm given below

```
Algorithm Name : Data Analysis using the C5.0 Mining Algorithm

Input: Combined Rule Pool $\mathcal{R}_{pool}$
Output: Data mining analysis results $Cl\_DM\_Rlst_n^{C5.0}$

Initialize $\mathcal{R}_n^{C5.0} = \emptyset$
Initialize $\mathcal{R}_{cnt} = 0$
For each $Rl_n \in \mathcal{R}_{pool}$
    For each $rl_{1mptyn} \in Rl_n$
        $\mathcal{R}_{cnt} = \mathcal{R}_{cnt} + 1$
        $\mathcal{R}_{ntmp}^{C5.0} = f_{rlmrg}^{C5.0}(rl_{1mptyn}, n)$
    End For
        $\mathcal{R}_n^{C5.0} = \mathcal{R}_n^{C5.0} + \mathcal{R}_{ntmp}^{C5.0}$
End For
Compute $Cl\_DM\_Rlst_n^{C5.0} = f_{clss}^{C5.0}(\mathcal{R}_n^{C5.0}, ddt_n)$
```

The proposed model eliminates the communication overheads arising from key exchange and at the same time preserves privacy as no party involved in the PPDM model exchanges any key amongst themselves owing to the adoption of Commutative RSA algorithm in our model. In the case of malicious parties involved in the system there is no notable data risk, as the data transacted over the network is in the cryptic form and is computationally indistinguishable proved in the subsequent sections of this paper. The next section of the paper discusses the computational complexity of the $KDLPPDM$ model constructed using a tri step approach.

7. COMPUTATIONAL ANALYSIS AND COMPARISONS

The amount of resources utilized in solving a computational algorithm or problem desired is known as the computational analysis. In this section of the paper the computational analysis of our proposed $KDLPPDM$ model is discussed. Furthermore this section discusses the computational efficiency of the $KDLPPDM$ model over the existing secure group communication models namely Secure Lock and Access Control Polynomial. The computational efficiency is measured in terms of the time complexity involved in solving the provided protocol on a homogenous computing environment. For comparisons the initialization step of the protocols is considered. In Secure Lock [6] and the Access Control Polynomial [7][8] a central server is considered for computing the cryptographic keys utilized for secure communication amongst the n parties considered. The central server computes the keys and distributes them securely to the parties involved in communication. The $KDLPPDM$ model does not consider a external or third party secure server to overcome the drawback of key distribution and key compromise attacks.

The PPDM initialization step discussed in the section six of this paper is responsible for the key computation and key derivation of the $KDLPPDM$ model. The computational complexity involved depend on the number of parties involved and the data transacted for key establishment and initialization. To prove that the $KDLPPDM$ model proposed through this paper is computationally less expensive when compared to the Secure Lock and Access Control Polynomial model we shall consider x number of parties involved in communication and each identified by an identity represented as $\mathcal{pty}_x$. The proposed Secure Lock communication protocol is based on the public and private key cryptography. Access Control Polynomial is based on symmetric key cryptography.

The $KDLPPDM$ model utilizes the benefits of Commutative RSA cryptography protocol for secure communication. Let us consider $\mathcal{K}ey_{S\_L}$ represents the private public key pairs applicable to the secure lock protocol. Let $\mathcal{K}ey_{A\_P}$ represents the symmetric key applicable to the Access Control Polynomial protocol and $\mathcal{K}ey_{PPDM}$ represent the encryption decryption key pairs applicable to the $KDLPPDM$ model. It is evident that greater the key size greater is the computation cost involved. A 1024 bit asymmetric cryptography is similar to an 80 bit symmetric key cryptography scheme [47] hence we consider the key lengths of $\mathcal{K}ey_{S\_L}$ and $\mathcal{K}ey_{PPDM}$ of 1024 bits in length and the key lengths of $\mathcal{K}ey_{A\_P}$ of 80 bit to use for comparative analysis. Let the computation function be represented by $\mathcal{O}_{com}$. Let's consider the Initialization, Key Computation and Derivation, Group Membership Verification, Encryption/Decryption Cost and storage cost at the server for comparisons.

The computational costs observed are as shown in the Table2 provided below.

Table 2. Algorithm based Computational Cost Analysis

PHASE OF ALGORITHM	SECURE LOCK	ACCESS CONTROL POLYNOMIAL	$KDLPPDM$
Initialization	$\mathcal{O}_{com}\left(x^2\mathcal{K}ey_{S\_L}^2\right) + \mathcal{O}_{com}(x\mathcal{K}ey_{S\_L}^3)$	$\mathcal{O}_{com}\left(x^2\mathcal{K}ey_{A\_P}^2\right)$	$\mathcal{O}_{com}(2x) + \mathcal{O}_{com}(2p\mathcal{K}ey_{PPDM}^2)$
Key Computation and Derivation	$\mathcal{O}_{com}\left(x\mathcal{K}ey_{S\_L}^2\right) + \mathcal{O}_{com}(\mathcal{K}ey_{S\_L}^3)$	$\mathcal{O}_{com}\left(x\mathcal{K}ey_{A\_P}^2\right)$	$\mathcal{O}_{com}(x\mathcal{K}ey_{PPDM}^2)$
Group Membership Verification	$\mathcal{O}_{com}(x\mathcal{K}ey_{S\_L})$	$\mathcal{O}_{com}(x\mathcal{K}ey_{A\_P})$	$Not\ Applicable$
Encryption/Decryption	$\mathcal{O}_{com}(x\mathcal{K}ey_{S\_L})$	$\mathcal{O}_{com}(x\mathcal{K}ey_{A\_P})$	$\mathcal{O}_{com}(x\mathcal{K}ey_{PPDM})$
Server End Storage	$\mathcal{O}_{com}(x + \mathcal{p}ty_x)$	$\mathcal{O}_{com}(x + \mathcal{p}ty_x)$	$Not\ Applicable$

All the above discussed models have been developed on the Visual Studio 2010 platform. The implementations were carried out using C#.Net. From Table 1 it is clear that the $KDLPPDM$ model does not consider the Group Membership Verification Phase and the Server End Storage computational overhead as there is no server involved. For comparisons the number of parties involved were varied form 10,20,30,50 70 and 100. The computational complexity of the Initialization Phase and the Key Computation and Derivation Phase of all the 3 models were measured in terms of the time involved in computation. The implementations were tested on an Intel Core 2 Duo 2.40 GHz CPU having 4GB 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.

The results obtained are as shown in Figure 1 it is evident that the Secure Lock protocol is computationally more expensive than the Access 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. The $KDLPPDM$ described in this paper does not consider a central trusted server for group establishment and cryptographic key distribution. In case of any malicious data custodians the data exchanged using Secure Lock and Access 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.

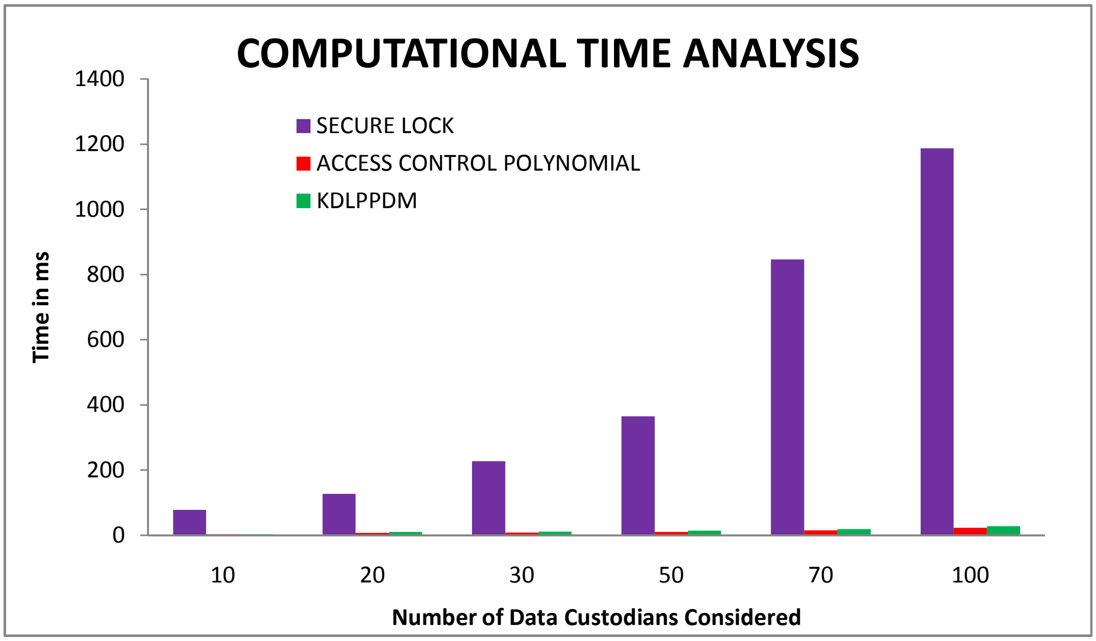

Figure 1. Computational Time Analysis

8. COMPUTATIONALLY INDISTINGUISHABLE ANALYSIS AND COMMUTATIVE NATURE PROOF OF KDLPPDM

The proof of privacy preserving provisioning is provided based on the computationally indistinguishablity [9][10][11] analysis of the data distribution observed in the proposed PPDM protocol. To demonstrate the proof of computational indistinguishablity a prototype PPDM system based on the $KDLPPDM$ was developed using the Visual Studio 2010 platform on the Microsoft .Net framework 4.0. The prototype implementation was considered assuming that the number of parties involved in the semi honest trust model is 3. The data set considered for analysis was the hyperthyroid data set. The hyperthyroid data set was vertically partitioned amongst the 3 parties considered. The C5.0 algorithm was utilized to generate the local rules at each parties end based on the vertical partitioned data each party housed. The C5.0 algorithm was run on an Linux Platform [31]. The commutative RSA algorithm was initialized in accordance to the $KDLPPDM$ Initialization algorithm discussed in the sixth section of the paper. On completion of the protocol initialization the locally generated rules were encrypted at each parties end and the creation of the secure rule pool was considered using Step 1 discussed earlier. The data distribution graphs of the transmissions post the secure rule pool creation is considered to provide the computational indistinguishablity proof. The data distribution frequency versus the data size graphs obtained are as shown

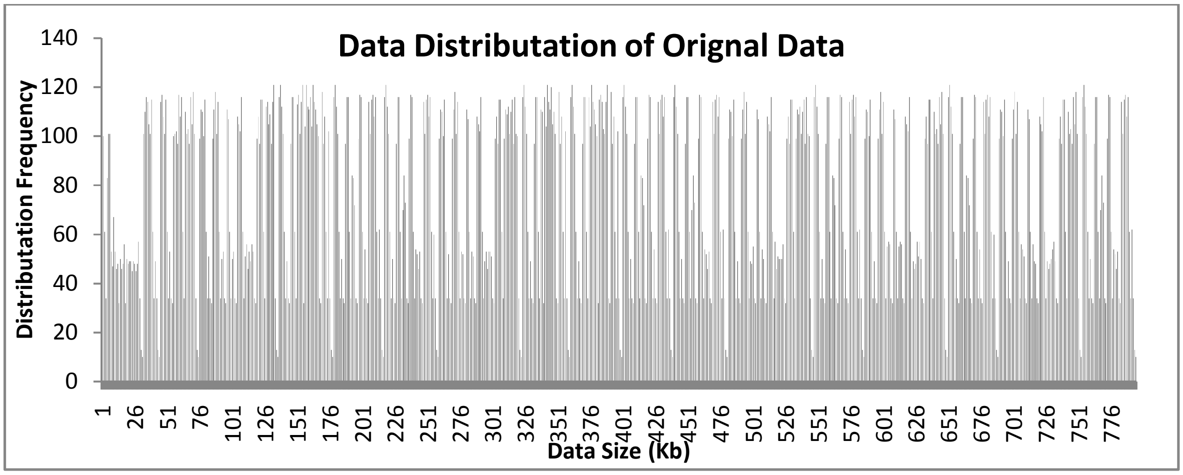

Figure 1. Data Distribution of Locally Generated Data Mining Rules for Data Custodian *P*

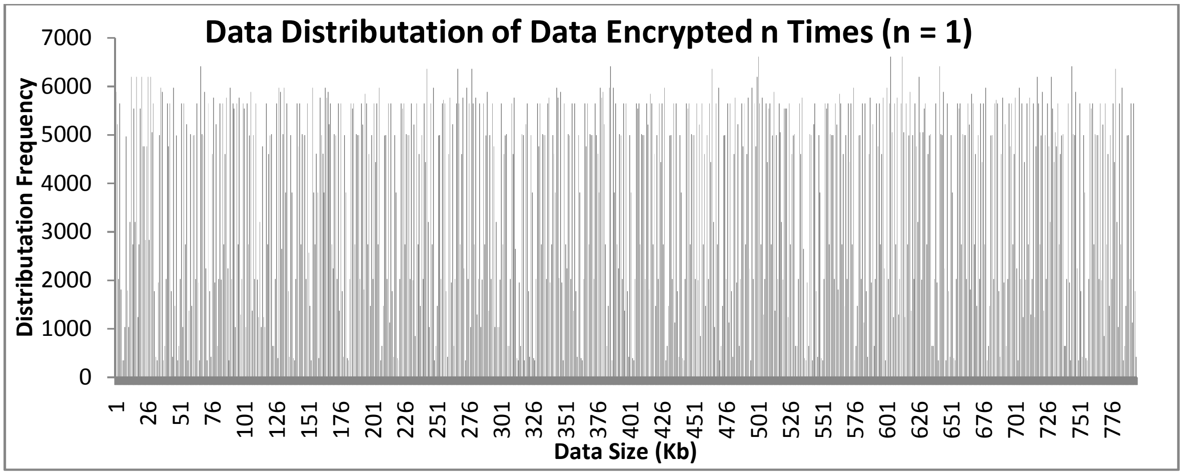

Figure 2. Data Distribution of Data Custodian aP's rules Encrypted n times using Commutative RSA $(n = 1)$

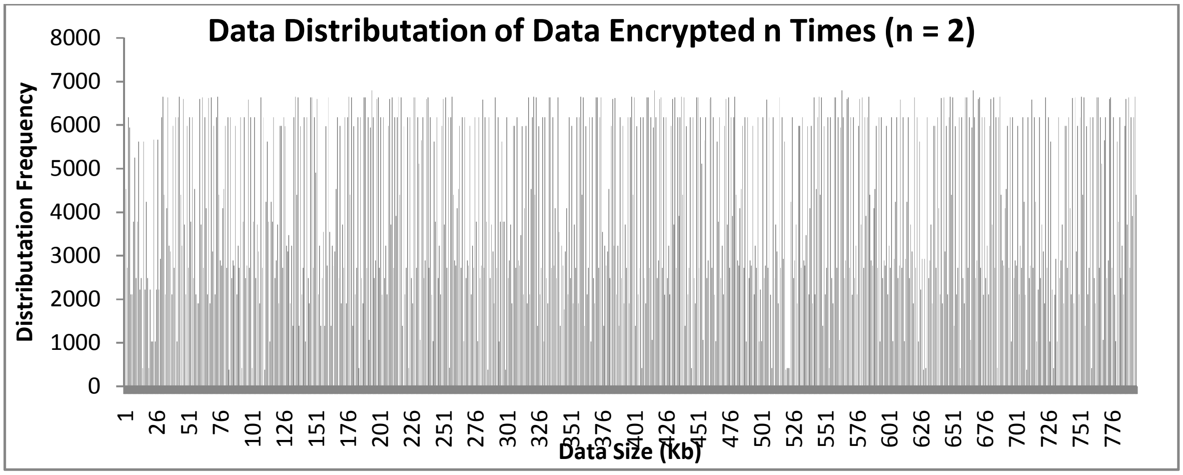

Figure 3. Data Distribution of Data Custodian P's rules Encrypted n times using Commutative RSA $(n = 2)$

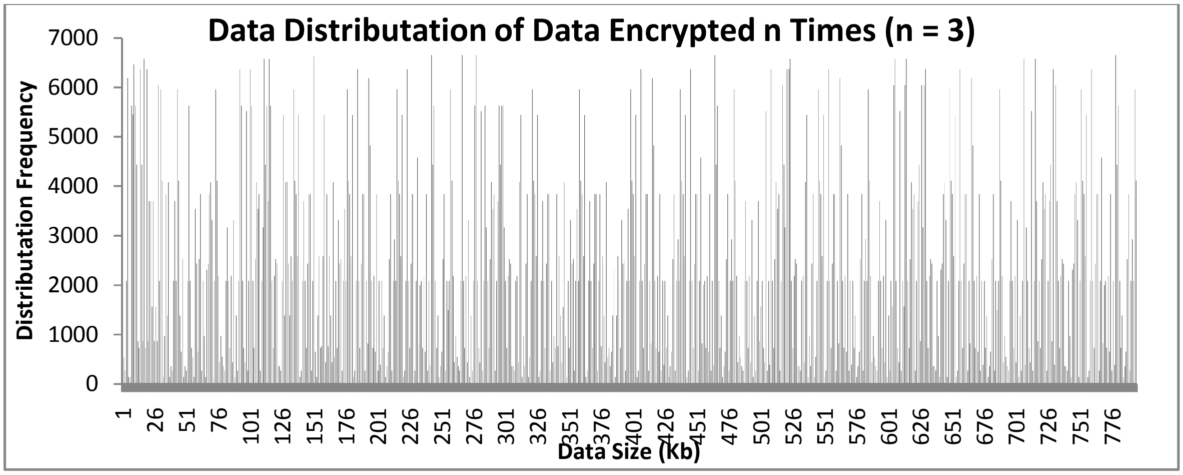

Figure 4. Data Distribution of Data Custodian P's rules Encrypted n times using Commutative RSA $(n = 3)$

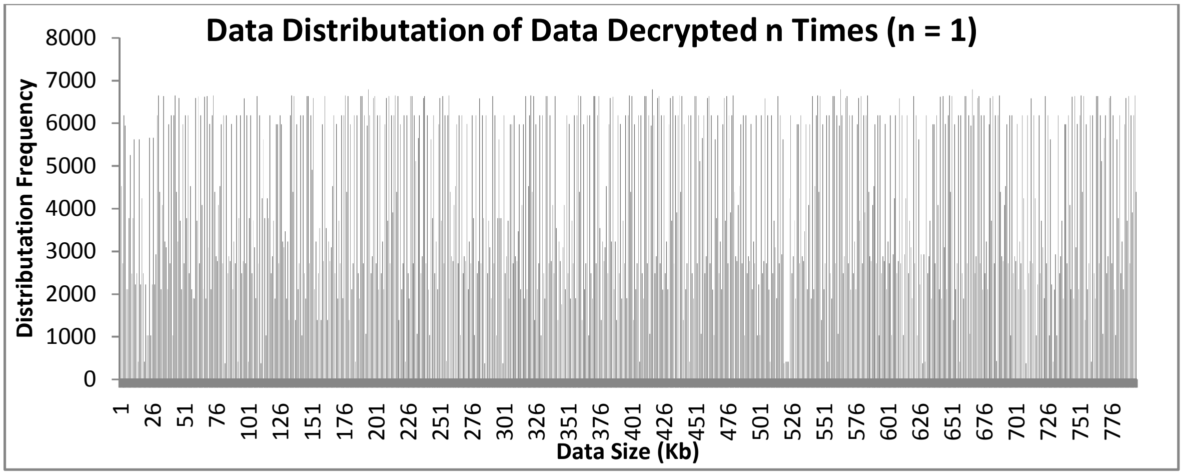

Figure 5. Data Distribution of Data Custodian P's rules Decrypted n times using Commutative RSA $(n = 1)$

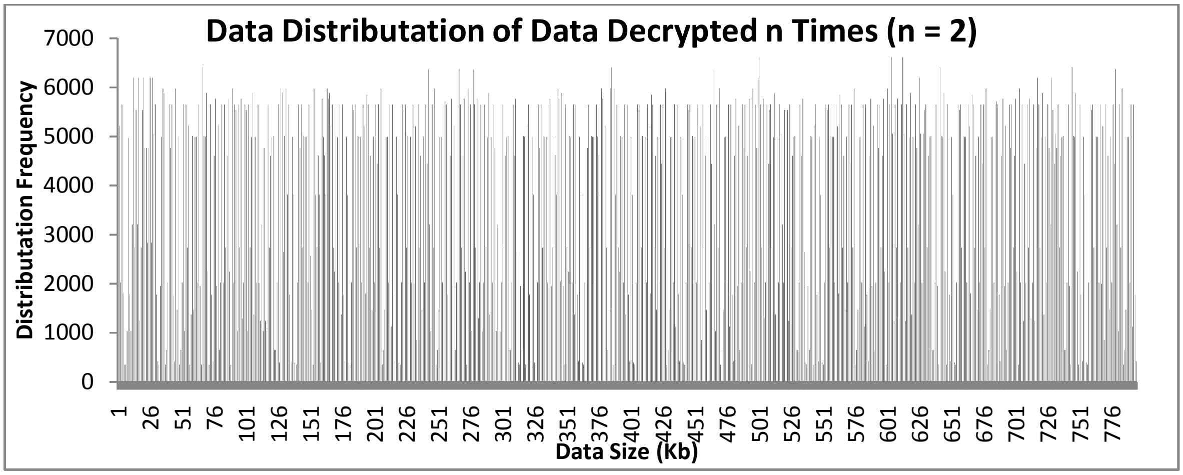

Figure 6. Data Distribution of Data Custodian P's rules Decrypted n times using Commutative RSA $(n = 2)$

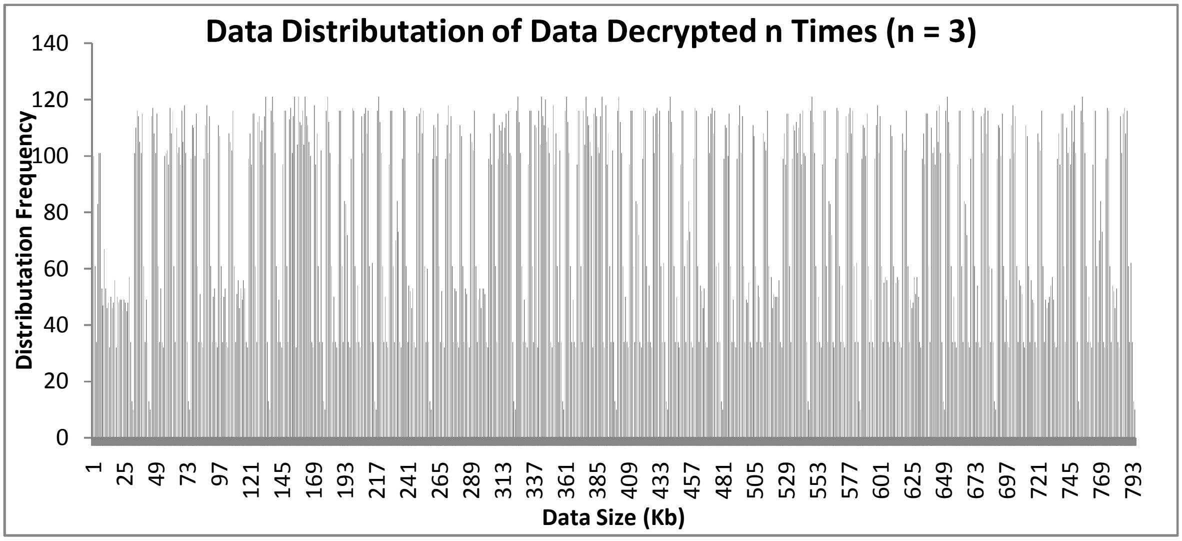

Figure 7. Data Distribution of Data Custodian $\boldsymbol{P}$'s rules Decrypted $\boldsymbol{n}$ times using Commutative RSA $(\boldsymbol{n} = \mathbf{3})$

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 8 , Figure 3 is identical to Figure 7 and Figure 4 is identical to Figure 6 which proves the commutative nature of the proposed system.

9. CONCLUSIONS AND FUTURE WORK

PPDM is utilized to derive knowledge when the data is housed in a distributed environment. This paper introduces the Key Distribution-Less Privacy Preserving Data Mining ($KDLPPDM$) model to preserve the privacy and provides an environment to attain desired knowledge extraction when the data available is vertically partitioned and distributed amongst the various parties involved. It is assumed that all the parties involved unite under a semi-honest trust model to achieve their respective mining goals. The $KDLPPDM$ introduced adopts the commutative RSA cryptographic algorithm to secure the data and preserve its privacy. The commutative property of the RSA algorithm eliminates the overheads and the risks involved in key computation, key distribution, key storage and key exchange even in the presence of colluded parties. The parties are reluctant to share the original local data amongst one and other, hence the proposed $KDLPPDM$ beleives in sharing the locally generated data mining rules providing for enhanced privacy preserving features. The use of the C5.0 data mining algorithm is adopted in the $KDLPPDM$to achieve higher mining accuracy, higher speed and for rule generation. The experimental evaluation presented proves that the proposed $KDLPPDM$ model and the Access Control Polynomial reduce the computational complexity by about 95.6% when compared to the Secure Lock mechanism. The privacy preserving feature of the $KDLPPDM$ is proved by the computational indistinguishablity analysis provided. The future of the research work presented in this paper is targeted towards an improved in depth understanding the C5.0 algorithm and to prove its efficiency over the existing data mining algorithms.

REFERENCES

[1] N.R. Adam and J.C. Wortmann, (1989) “Security-Control Methods for Statistical Databases: A Comparison Study”, *ACM Computing Surveys*, Vol. 21, No. 4, pp. 515-556.

[2] D.G. Marks, (1996) “Inference in MLS Database,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 1, pp. 46-55.

[3] M.-S. Chen, J. Han, and P. S. Yu, (1996) “Data mining: An overview from database perspective,” *IEEE Transactions on Knowledge and. Data Engineering*, Vol. 8, No. 6, pp. 866–883.

[4] C. Clifton and D. Marks, (1996) “Security and Privacy Implications of Data Mining”, *Proc. 1996 ACM SIGMOD Int'l Workshop Data Mining and Knowledge Discovery*, pp. 15-16.

[5] Matthews, Gregory J., Harel, Ofer, (2011) “Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy”, Statistics Surveys, Vol. 5, pp. 1-29.

[6] G. H. Chiou and W. Chen, (1989) “Secure broadcasting using the Secure Lock”, *IEEE Transactions on Software Engineering*, Vol. 15, No. 8, pp. 929–934.

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

[8] Xukai Zou, Mingrui Qi, and Yan Sui,(2011) “A New Scheme for Anonymous Secure Group Communication”, *System Sciences (HICSS) 44th Hawaii International Conference.*

[9] Oded Goldreich, (2001) “*Foundations of Cryptography: Basic Tools*”, Vol. 1. Cambridge University Press.

[10] Oded Goldreich, (2004) “*Foundations of Cryptography: Basic Applications*”, Vol. 2. Cambridge University Press.

[11] Alexandre Evfimievski and Tyrone Grandison, (2009) "Privacy-Preserving Data Mining", *Encyclopedia of Database Technologies and Applications*, IGI Global, pp. 1-8.

[12] V.S. Verykios et al., (2004) “State-of-the-Art in Privacy Preserving Data Mining”, *SIGMOD Record,* Vol. 3, No. 1, pp. 50-57.

[13] C. Clifton et al., (2003) “Tools for Privacy Preserving Distributed Data mining”, *SIGKDD Explorations*, Vol. 4, No. 2, pp. 28-34.

[14] Nan Zhang and Wei Zhao, (2007) "Privacy-Preserving Data Mining Systems" ,*IEEE Computer*, Vol. 40, No. 4, pp. 52-58.

[15] Y. Lindell and B. Pinkas, (2000) “Privacy preserving data mining”, *Springer-Verlag Advances in Cryptology,* pp. 36-53.

[16] *R. Agrawal and R. Srikant, (2000)* “Privacy preserving data mining” *In Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 439-450.

[17] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang, (2012) “Enabling Multilevel Trust in Privacy Preserving Data Mining”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 9, pp. 1598–1612.

[18] Ming-Jun Xiao, Kai Han, Liu-Sheng Huang and Jing-Yuan Li, (2006) "Privacy Preserving C4.5 Algorithm Over Horizontally Partitioned Data", *Proceedings of the Fifth IEEE International Conference on Grid and Cooperative Computing,* pp. 78-85.

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

[20] A.W.-C. Fu, R. C.-W.Wong, and K.Wang, (2005) “Privacy-preserving frequent pattern mining across private databases”, *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 613-616.

[21] J. Vaidya and C. W. Clifton, (2002) “Privacy preserving association rule mining in vertically partitioned data,” *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data* mining (KDD’02), pp. 639-644.

[22] D. Agrawal and C. C. Aggarwal, (2001) “On the design and quantification of privacy preserving data mining algorithms”, *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '01)*, pp. 247-255.

[23] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, (2002) “Privacy preserving mining of association rules”, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD '02)*, pp. 217-228.

[24] K. Chen and L. Liu, (2005) “Privacy preserving data classification with rotation perturbation”, *Fifth IEEE International Conference on Data Mining.*

[25] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, (2007) “Time series compressibility and privacy”, *Proceedings of the 33rd VLDB International Conference on Very Large Databases (VLDB '07), Vienna*, Austria.

[26] L.Sweeney, (2002) “k-anonymity: A model for protecting privacy”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)*, Vol. 10, No. 5, pp. 557-570.

[27] C. C. Aggarwal and P. S. Yu, (2004) “A condensation approach to privacy preserving data mining,” *Proceedings of the International Conference on Extending Database Technology (EDBT)*, vol. 2992, pp. 183-199.

[28] Slava Kisilevich, Lior Rokach, Yuval Elovici and Bracha Shapira, (2010) "Efficient Multidimensional Suppression for K-Anonymity", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 3, pp. 334-347.

[29] W. Du and Z. Zhan, (2003) “Using randomized response techniques for privacy-preserving data mining”, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 505-510.

[30] R. Agrawal, R. Srikant, and D. Thomas, (2005) “Privacy preserving OLAP”, *Proceedings of the. ACM SIGMOD International Conference. On Management of Data,* pp. 251-262.

[31] http://www.rulequest.com/see5-info.html

[32] 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. (2008) "Top 10 algorithms in data mining", *Springer, Knowledge and Information Systems* Vol. 14, No. 1, pp. 1-37.

[33] Tomasz Bujlow, Tahir Riaz, Jens Myrup Pedersen, (2012) "A method for classification of network traffic based on C5.0 Machine Learning Algorithm", *in Workshop on Computing, Networking and Communications, IEEE*, pp 237-241.

[34] Meng Wang, Kun Gao , Li-jing Wang and Xiang-hu Miu, (2012) “A Novel Hyperspectral Classification Method Based on C5.0 Decision Tree of Multiple Combined Classifiers”, *Fourth International Conference on Computational and Information Sciences (ICCIS)*, pp. 373- 376.

[35] Yanguang Shen, Hui Shao and Li Yang, (2009) "Privacy Preserving C4.5 Algorithm over Vertically Distributed Datasets", *IEEE, International Conference on Networks Security, Wireless Communications and Trusted Computing*, Vol.2, pp. 446-448.

[36] Po-Hsun. Sung, Jyh-Dong Lin, Shih-Huang Chen, Shun-Hsing Chen, and Jr-Hung Peng, (2010) “Utilization of Data Mining on Asset Management of Freeway Flexible Pavement”, *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),* Vol.10, pp. 977 – 979.

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

[38] B. Pinkas, (2002)" Cryptographic techniques for privacy-preserving data mining", *ACM SIGKDD Explorations Newsletter*, Vol. 4, No. 2, pp. 12-19.

[39] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu "Tools for Privacy Preserving Distributed Data Mining",*SIGKDD Explorations,*Vol. 4, No. 2, pp 1-7.

[40] Victor P. Hubenko Jr., Richard A. Raines, Rusty O. Baldwin, Barry E. Mullins, Robert F. Mills, and Michael R. Grimaila,(2007) "Improving Satellite Multicast Security Scalability by Reducing Rekeying Requirements",*IEEE Network* ,Vol. 21, No.4, pp. 51-56.

[41] Bezawada Bruhadeshwar and Sandeep S. Kulkarni, (2011) "Balancing Revocation and Storage Trade-Offs in Secure Group Communication", *IEEE Transactions on Dependable and Secure Computing*, Vol. 8, No. 1, pp. 58-73.

[42] Nathaniel Karst, and Stephen B. Wicker, (2011) "On the Rekeying Load in Group Key Distributions Using Cover-Free Families ", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 58, No. 10, pp. 6667 – 6671.

[43] J. Vaidya and C. Clifton, (2003) "Privacy-preserving k-means clustering over vertically partitioned data," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data (KDD '03)*, pp. 206-215.

[44] Patrick Sharkey, Hongwei Tian, Weining Zhang, and Shouhuai Xu, (2008) "Privacy-Preserving Data Mining through Knowledge Model Sharing", *Springer-Verlag* ,Berlin Heidelberg 2008 LNCS 4890, pp. 97–115.

[45] Gabriel Ghinita, Panos Kalnis, and Yufei Tao, (2011) "Anonymous Publication of Sensitive Transactional Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 2,. Pp. 161-174.

[46] Pui K. Fong and Jens H. Weber-Jahnke, (2012) "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 2, pp. 353 – 364.

[47] Arjen K. Lenstra, "Key length. Handbook of Information Security", Editor-in-Chief, Hossein Bidgoli, pp. 617–635.

[48] Khatri Nishant P., Preeti Gupta and Tusal Patel, "Privacy Preserving Clustering on Centralized Data Through Scaling Transformation", *International Journal of Computer Engineering & Technology (IJCET),* Volume 4, Issue 3, 2013, pp. 449 - 454, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[49] D.Pratiba and Dr.G.Shobha, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing", *International Journal of Computer Engineering & Technology (IJCET),* Volume 4, Issue 3, 2013, pp. 441 - 448, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[50] Sumana M and Hareesha K S, "Preprocessing and Secure Computations for Privacy Preservation Data Mining", *International Journal of Computer Engineering & Technology (IJCET),* Volume 4, Issue 4, 2013, pp. 203 - 212, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[51] R. Manickam, D. Boominath and V. Bhuvaneswari,, "An Analysis of Data Mining: Past, Present and Future*", International Journal of Computer Engineering & Technology (IJCET)*, Volume 3, Issue 1, 2012, pp. 1 - 9, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

BIOGRAPHIES

Kumaraswamy S is currently working as an Assistant Professor in the Department of Computer Science and Engineering, KNS Institute of Technology, Bangalore, India. He obtained his Bachelor of Engineering from SiddaGanga Institute of Technology, Bangalore University, Tumkur. He received his M E Degree in Computer Science and Engineering from UVCE, Bangalore University, Bangalore. He is presently pursuing his Ph.D programme in the area of privacy management in databases in Bangalore University. His research interest is in the area of Data mining, Web mining and Semantic web.

S H Manjula is currently the Chairman, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. She obtained her Bachelor of Engineering and Masters Degree in Computer Science and Engineering from University Visvesvaraya College of Engineering (UVCE). She was awarded Ph.D in Computer Science from Dr. MGR University, Chennai. Her research interests are in the field of Wireless Sensor Networks and Data mining.

K R Venugopal is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 39 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc.. During his three decades of service at UVCE he has over 350 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

L M Patnaik is currently Honorary Professor, Indian Institute of Science, Bangalore, India. He was a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India and was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 500 research publications in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.