Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database

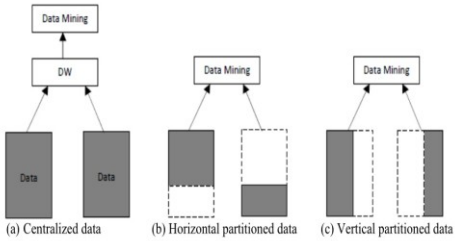
RACHIT V. ADHVARYU, NIKUNJ H. DOMADIYA

***Abstract* - The advances of data mining techniques played an important role in many areas for various applications. In context of privacy and security issues, the problems caused by association rule mining technique are recently investigated. The misuse of this technique may disclose the database owner's sensitive information to others. Hence, the privacy of individuals is not maintained. Many of the researchers have recently made an effort to preserve privacy of sensitive knowledge or information in a real database. In this paper, we have modified EMHS Algorithm to improve its efficiency by using Elliptic Curve Cryptography. We have used ElGamal Cryptography technique of ECC for homomorphic encryption. Analysis of the experiment on various datasets show that proposed algorithm is efficient compared to EMHS in terms of computation time.**

Keywords: **Data Mining, Elliptic Curve Cryptography, ElGamal Cryptography, EMHS, Privacy, Privacy Preserving Association Rule Mining**

I INTRODUCTION

Data mining or knowledge discovery techniques such as association rule mining, classification, clustering, sequence mining, etc. have been most widely used in today's information world [1]. Successful application of these techniques has been demonstrated in many areas like marketing, medical analysis, business, Bioinformatics, product control and some other areas that benefit commercial, social and humanitarian activities. These techniques have been demonstrated in centralized as well as distributed environments. In centralized environment, all the datasets are collected at central site (data warehouse) and then mining operation is performed, as shown in Fig (a), where in distributed environment, data may be distributed among different sites which are not allowed to send their data to find global mining result. There are two types of distributed data considered. One is horizontally partitioned data and another is vertically partitioned data. As shown in Fig. (b) And Fig. (c) Data are distributed among two sites which wish to find the global mining result. The horizontal partitioned data shown in Fig. (b) Where Fig. (c) Shows vertical partitioned data. In horizontal partitioned data, each site contains same set of attributes, but different number of transactions wherein vertical partitioned data each site contains different number of attributes but same number of transactions [1].



Different Database Environments

Recently these techniques are investigated in terms of privacy and security issues and it is concluded that these techniques threat to the privacy of individuals information. That means one (e.g. adversary or malicious user) can easily infer someone's sensitive information (or knowledge) by mining technique. So, sensitive information should be hidden in database before releasing. For distributed mining it should be protected from the involving parties (or sites) who wish to find global mining result[2]. Therefore, to preserve privacy for sensitive knowledge, privacy preserving data mining (PPDM) become a hot directive in data or knowledge engineering field.

II ASSOCIATION RULE MINING

Association Rule Mining is a popular technique in data mining for discovering interesting relations between items in large databases. It is purposeful to identify strong rules discovered in the databases using different available measures. Based on the concept of strong rules, Rakesh Agrawal et al [3]. described association rules for discovering similarities between products in large-scale transaction data in supermarkets. For example, the rule {Bread, Butter} =>{Milk} found in the sales data of a shop would indicate that if a customer buys bread and butter together, he or she is likely to also buy milk. Such information can be used in decision making about marketing policies such as, e.g., product offers, product sales and discount schemes. In addition to the above mentioned example association rules are used today in many application areas including Web usage mining, Intrusion detection, Continuous production, and Bioinformatics

ISSN: 2278 – 1323

[3]. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

The problem of association rule mining [3] is defined as: Let I= $\{i_1, i_2,\ldots, i_n\}$ be a set of n binary attributes called *items*. Let D=$\{t_1,t_2,\ldots,t_m\}$ be a set of transactions called the *database*. Each transaction in database D has a unique transaction identity ID and contains a subset of the items in I [3]. A *rule* is defined as an implication of the form X=>Y where X,Y is subset of I and X intersection Y = Null Set. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

Support count: The support count[3] of an itemset X, denoted by X.count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions. Then

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

The most famous application of association rules is its use for Market Basket Analysis [4]. Association Rules are helpful in many fields like Telecommunication and Medical records for retrieving some desired results. Association rules has been used in mining web server log files to discover the patterns that accesses different resources continuously or accessing particular resource at regular interval. Association rules are also useful in mining census data, text document, health insurance and catalog design [4].

III RELATED WORK

To understand the background of privacy preserving in association rule mining, we present different techniques and algorithm in the following subsections.

A SECURE MULTIPARTY COMPUTATION (SMC) WITH TRUSTED THIRD PARTY

This technique worked as a client server system where one site is a server responsible for the generating global result and all remaining sites are client sites which sends its encrypted data to the server to retrieve global result [5].

An example to SMC with trusted third party was PPDM-ARBSM algorithm [6]. This algorithm has mainly two servers: Data Mining Server and Cryptosystem Management Server [6]. A disadvantage of this algorithm was that the failure of third party fails the communication.

B SECURE MULTIPARTY COMPUTATION (SMC) WITH SEMI HONEST MODEL

This technique assumes all the sites as honest. One site acts as an initiator [7] and all others as sites. All the sites send their encrypted data to the next site in queue. Finally the last site sends all data to initiator which finds the global result [7].

An example to SMC with semi honest model was Fast Private Association Rule Mining for Securely Sharing algorithm [8]. The detailed description is mentioned in [8]. The limitation of this algorithm was the increase in computation time with the increase in the number of sites.

C MHS ALGORITHM FOR SECURE SHARING ON HORIZONTALLY PARTITIONED DATABASE.

MHS algorithm worked on minimum 3 sites. One site acts as an Initiator, one site acts as Combiner [9]. This algorithm used RSA cryptosystem. All sites find its frequent itemsets, encrypt it using RSA public key and send it to Combiner. The task of the combiner is to merge all the data with its own data and send it to the initiator. The task of the initiated was to decrypt all the data and generate global results [9]. As this algorithm was based on the concept of frequent itemsets, the limitation was the increase in computation time with the increase in the database size and number of sites.

D EMHS ALGORITHM FOR PRIVACY PRESERVING ASSOCIATION RULE MINING ON HORIZONTALLY PARTITIONED DATABASE.

To implement EMHS algorithm, minimum 3 sites are required. One site acts as an Initiator, the other acts as Combiner and rest all are client sites. The algorithm was based on the following concepts as mentioned below.

A. MFI (Maximal Frequent Itemset):
A Frequent Itemset which is not a subset of any other frequent itemset is called MFI. By using MFI, communication cost is reduced [10]. FI was replaced by MFI

B. RSA Algorithm:
one of the widely used public key cryptosystem. It is based on keeping

ISSN: 2278 – 1323

factoring product of two large prime numbers secret. Breaking RSA encryption is tough [10]. This was used in the first phase.

C. *Homomorphic Paillier Cryptosystem:*

Paillier cryptosystem is an additive homomorphic cryptosystem, meaning that one can compute cipher texts into a new cipher text that is encryption of sum of the messages of the original cipher texts. For E.g. Let m1, m2 be the two messages. Then Encryption= E (m1+m2) =E (m1) *E (m2) and Decryption= D (E (m1) *E (m2)) =m1+m2 i.e. the sum of m1 and m2. Also, if the size of the public key is to (bit) then the size of cipher text c is 2*t (byte) [10].

EMHS algorithm was implemented in 3 phases. In the first phase, RSA cryptosystem was used. While in the second and third phase Homomorphic Paillier cryptosystem was used. The results showed better performance in the mining process as compared to other algorithms.

IV PROPOSED NEW ALGORITHM

A BASIC CONCEPTS OF NEW ALGORITHM

Suppose database D is distributed among n sites (S1,S2,..,Sn) in such a way that database Di ($1 \leq i \leq n$) containing site Si consists of same set of attributes but different number of transactions. All sites are considered as semi honest. Now the problem is to mine valid global association rules satisfying given minimum support threshold (MST) and minimum confidence threshold (MCT) in unsecured environment, which should fulfill following privacy and security issues.

1) No any involving party should be able to know the contents of the transaction of any other involving parties.

2) Adversaries should not be able to affect the privacy and security of the information of involving parties by reading communication channel between involving parties.

B ELLIPTIC CURVE CRYPTOGRAPHY

Elliptic curve provides public cryptosystem based on the discrete logarithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide a security level same as RSA with larger key length. A detailed overview of elliptic curves and an elliptic curve cryptosystem is given in [11][12]. We used ElGamal Cryptography [13] in our proposed algorithm. In the following, we give an overview of ElGamal cryptography.

1. ELGAMAL CRYPTOGRAPHY

a) A wishes to exchange message M with B[13].
b) B first chooses Prime Number p, Generator g and private key x.
c) B computes its Public Key $Y = g^x \bmod p$ and sends it to A.
d) Now A chooses a random number k.
e) A calculates one time key $K = Y^k \bmod p$.
f) A calculates $C1 = g^k \bmod p$ and $C2 = M*K \bmod p$ and sends (C1,C2) to B.
g) B calculates $K = C1^x \bmod p$
h) B calculates K^{-1} = inverse of K mod p
i) B recovers $M = K^{-1} * C2 \bmod p$
j) Thus, Message M is exchanged between A and B securely[13]

2. ELGAMAL EXAMPLE

- A wishes to exchange message 100 with B.
- B first chooses Prime Number p = 139, Generator g = 3 and private key x = 12.
- B computes its Public Key $Y = 3^{12} \bmod 139 = 44$ and sends it to A.
- Now A chooses a random number k = 52.
- A calculates one time key $K = 44^{52} \bmod 139 = 112$.
- A calculates $C1 = 3^{52} \bmod 139 = 38$ and $C2 = 100*112 \bmod 139 = 80$ and sends (38,80) to B.
- B calculates $K = 38^{12} \bmod 139 = 112$ (same as one time key of A)
- B calculates $K^{-1} = 112^{-1} \bmod 139 = 36$
- B recovers $M = 36 * 80 \bmod 139 = 100$ (M)

C PROPOSED COMMUNICATION PROTOCOL

The proposed communication protocol is defined in Fig. 2. Suppose there are 3 sites, namely SITE 1, SITE 2, SITE 3. Among them, there are 2 sites, namely Initiator and Combiner. All the parties are semi-honest. Suppose that they want to find the global results without revealing their information to other sites.

The proposed communication protocol is same as EMHS algorithm. But we use ElGamal Cryptography

ISSN: 2278 – 1323

instead of RSA and Homomorphic Paillier Cryptosystem.

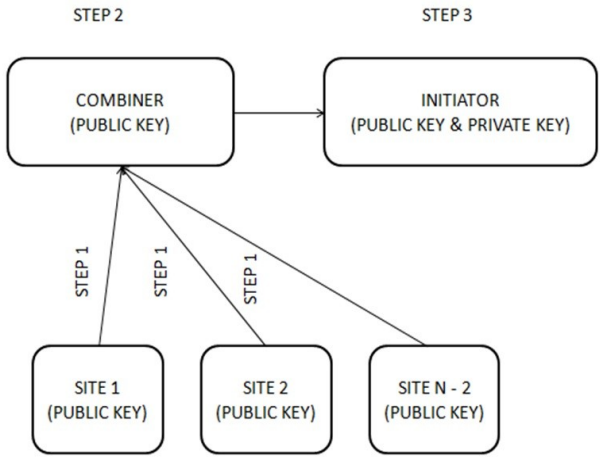


Fig 2. Proposed Communication Protocol.

We first how to use Paillier cryptosystem to compute global support counts.

Lemma 1: With itemset X and (n-1) sites, the global support counts can be done as follows:

Encryption: $E(X_{.sup1} + \ldots + X_{.sup(n-1)}) = E(X_{.sup1}) * \ldots * E(X_{.sup(n-1)})$.

Decryption: $D(E(X_{.sup1}) * \ldots * E(X_{.sup(n-1)})) = X_{.sup1} + \ldots + X_{.sup(n-1)}$.

After decryption, the result is the sum of support counts of X at sites (n-1).

The detailed description of our proposed algorithm is as follows:

Phase 1:

a) The initiator generates ElGamal and Paillier public key and private key. It sends public keys to combiner and all other client sites.
b) Each site, except initiator computes its MFI, encrypts it using ElGamal public key and sends it to the Combiner.
c) The combiner merges the received data with its own data and sends it to the Initiator.
d) Initiator decrypts the received data using ElGamal private key. Then it adds its own data with the decrypted data and computes to find global MFI. Then the result is sent to all other sites.

Phase 2:

a) Each site finds frequent itemset and its local support count on the basis of MFI.
b) Each site, except initiator encrypts the data using Paillier's public key and sends it to the Combiner.
c) Combiner merges its own data with the received data. After this, encrypted data is sent to Initiator.
d) Initiator decrypts the received data using Paillier private key. It generates a global support count of each candidate X as:
$X.sup = D(E(X.combiner)) + X.sup_{Initiator}$

Phase 3:

a) Each site finds its database size |DB|.
b) Each site, except initiator encrypts the data using Homomorphic Paillier public key and sends it to the combiner.
c) The combiner merges its own data with the received data and sends it to the initiator.
d) Initiator decrypts the received data using Paillier private key. Then it adds its own data with the decrypted data and computes to find global database size |DBi|.
e) Finally, Initiator generates the global association rules and sends the result to all other sites.

Our proposed system does not reduce the steps of communication, but due to smaller key size, it will reduce the computation time and increase the performance.

V DISCUSS AND EXPERIMENTAL RESULTS

In this section, we discuss improved EMHS basing on the criteria of performance in terms of computation time. The comparison of EMHS and Proposed algorithm is presented in detail.

Firstly, we discuss about the privacy. EMHS and our algorithm both satisfies semi-honest model. The smaller key size of ECC provides equivalent security as compared to RSA. Thus the privacy remains the same in EMHS and proposed algorithm.

Lastly, the discussion is about the performance in which computation time is the main measure. Generally the proposed algorithm improves the performance in phases when increases the number of sites. The detailed comparison of performance between EMHS and the proposed algorithm on real datasets is described as follows.

ISSN: 2278 – 1323

Both EMHS and Newly Proposed System are executed with the number of sites, increasing from 3 to 7 on four real datasets: Chess, Connect, Mushroom, Pumsb. All these datasets have different features as explained in Table 1.

In implementation, each dataset is divided into 3 to 7 parts on the basis of the records. **Frequent Pattern Mining Framework (FPMF)** is used to implement proposed algorithm. **CharmMFI** algorithm is used to find local MFI at each site. **Eclat** algorithm is used to find local frequent itemsets from the global itemsets at each site in the proposed system.

DATABASE	ROWS	COLUMNS	DENSE / PARSE?
CHESS	3196	37	DENSE
CONNECT	5000	46	DENSE
MUSHROOM	8124	23	DENSE
PUMSB	6000	76	DENSE

Table 1. Description of Database

The global support count for each dataset is 80% (0.80) respectively.

Fig. 3, Fig. 4, Fig. 5, and Fig. 6 shows the comparison results based on performance of EMHS and proposed algorithm on the entire above mentioned databases.

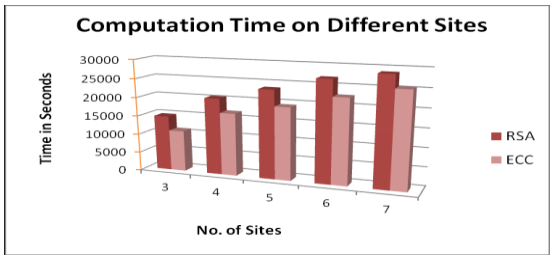


Fig. 3 Comparison on CHESS

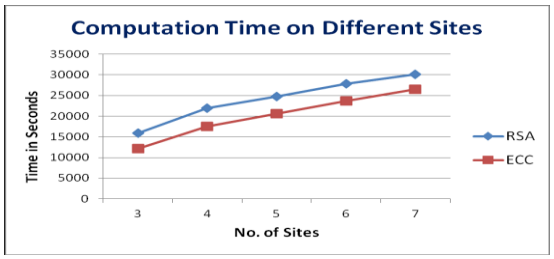


Fig. 4 Comparison on CONNECT

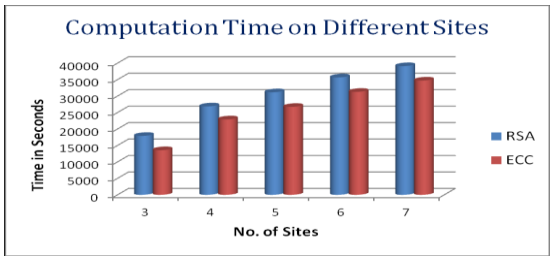


Fig. 5 Comparison on MUSHROOM

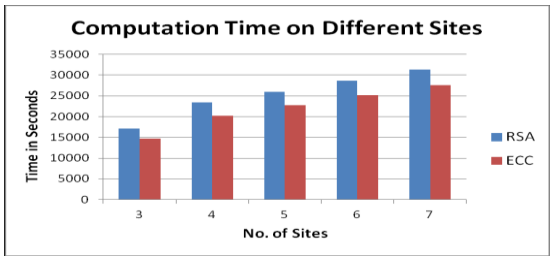


Fig. 6 Comparison on PUMSB

VI CONCLUSION

In this paper, we proposed an algorithm to improve privacy and performance of EMHS when increasing the number of sites. We maintain the model of EMHS and apply ElGamal Cryptography in the first phase and Paillier cryptosystem in the second phase.

From the experimental results we conclude that the proposed system has better performance than EMHS in dense datasets when increasing the number of sites. In future work, the collusion of Initiator and Combiner must be solved.

REFERENCES

[1] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, "Disclosure limitation of sensitive rules," in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX " 99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 45–52

[2] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," in National Science Foundation Workshop on Next Generation Data Mining, 2002, pp. 126–133.

[3] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000)", 439–450.

[4] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis: "State-of-the-art in Privacy Preserving Data Mining", March 2004.

[5] N V Muthu Lakshmi and Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining Without Trusted Party For Horizontally Partitioned Databases", International

ISSN: 2278 – 1323

Journal of Data Mining AND Knowledge Management Process (IJDKP) Vol.2, No.2 March 2012
[6] GUI Qiong, CHENG Xiao-hui, "A Privacy-Preserving Distributed Method for Mining Association Rules", 2009 International Conference on Artificial Intelligence and Computational Intelligence,pp294-297 2009.
[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques". San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 227-245.
[8] Estivill-Castro, V., Hajyasien, "A Fast Private Association Rule Mining by a Protocol Se-curely Sharing Distributed Data". In: Proceedings of the 2007 IEEE Intelligence and Secu-rity Informatics (ISI 2007), New Brunswick, New Jersey, USA, May 23-24, pp. 324-330 2007
[9] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail, "Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base", Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607-616 2008.
[10] Xuan C. N., Hoai B. L., Tung A. C., "An enhanced scheme for privacy preserving association rules mining on horizontally distributed databases, 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF"), pp 1 - 4 2012.
[11] N. Koblitz, "Elliptic curve cryptosystems," Mathematics of Computation, vol. 48, pp. 203–209, 1987.
[12] M. Anoop, "Elliptic curve cryptography."
[13] William Stallings, "Cryptography and Network Security", Fifth Edition, 2011

Rachit Adhvaryu is a student of Masters of Engineering in Computer Science & Engineering at B. H. Gardi College of Engineering and Technology, Rajkot, Gujarat, India. He is bachelors in Computer Engineering. His area of interest are Data Mining, Database and Security

Nikunj Domadiya is an Assistant Professor at B. H. Gardi College of Engineering and Technology, Rajkot, Gujarat, India. He is Bachelors and Masters in Computer Engineering. His area of interests are Data Mining, Information Security, Cryptography etc.

ISSN: 2278 – 1323