Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database

[1]Manvar Sagar R., [2]Desai Sonali P., [3]Hathi Karishma B.
[1,23,] PG Student CE Department,
B.H.Gardi College of Engg.&Tech., Rajkot, Gujarat India

Abstract: **The advancement in data mining techniques plays an important role in many applications. In context of privacy and security issues, the problems caused by association rule mining technique are investigated by many research scholars. It is proved that the misuse of this technique may reveal the database owner's sensitive and private information to others. Many researchers have put their effort to preserve privacy in Association Rule Mining. Amongst the two basic approaches for privacy preserving data mining, viz. Randomization based and Cryptography based, the later provides high level of privacy but incurs higher computational as well as communication overhead. Hence, it is necessary to explore alternative techniques that improve the over-heads. In this work, we propose an efficient, collusion-resistant cryptography based approach for distributed Association Rule mining using Shamir's secret shar-ing scheme. As we show from theoretical and practical analysis, our approach is provably secure and require only one time a trusted third party. We use secret sharing for privately sharing the information and code based identification scheme to add support against malicious adversaries.**

Keywords**: Privacy, Privacy Preservation In Data Mining (PPDM), Horizontally Partitioned Database, EMHS,MFI, Shamir secret Sharing.**

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].

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 [2].

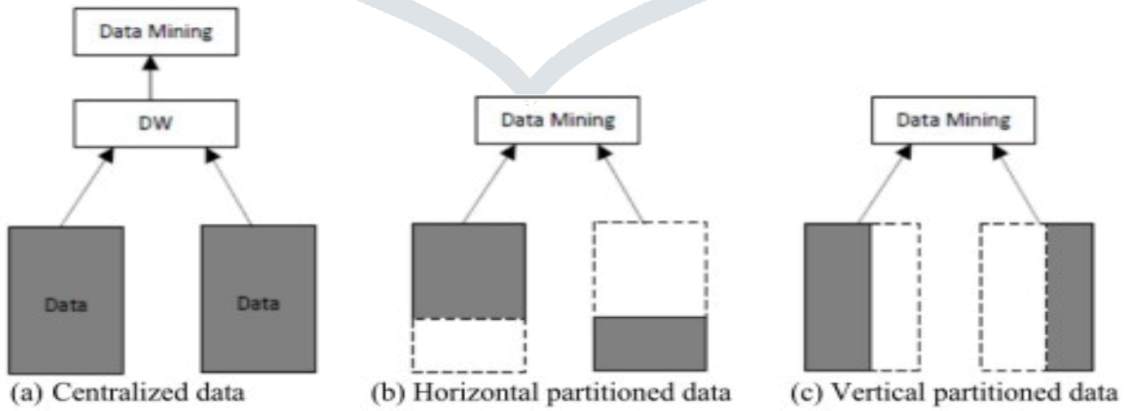


Figure 1.Different Database Environments

Privacy-preserving data mining (PPDM) has been studied extensively and applied widely. In this paper, we are particularly interested in the mining of association rule in a scenario where the data is horizontally distributed among different parties .To mine the association rule, these parties need to collaborate with each other so that they can jointly mine the data and produce results that interest all of them. However, Due to privacy law and motivation of business interests, the parties may not trust any other parties and do not want to reveal her own portion of the data, although realizing that combining their data has some mutual benefit.

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 [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= {i1, i2,…, in} be a set of n binary attributes called *items*. Let D={t1,t2,…,tm} 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][15] 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

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

$$\text{Confidence} = \frac{(X \cup Y).\text{count}}{\text{c.count}}$$

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.

Goal: Find the global result of aggregate database using all local database with maintain Privacy and database quality [5].

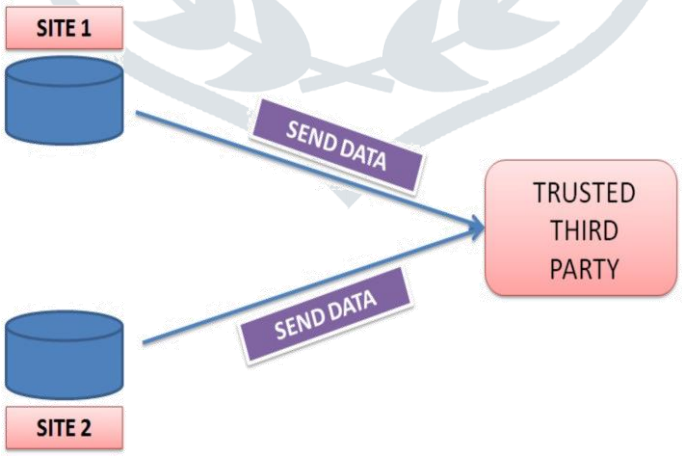


Figure 2: Distributed Scenario

The Detailed Scenario of Distributed Database is shown in Fig. 3[6]

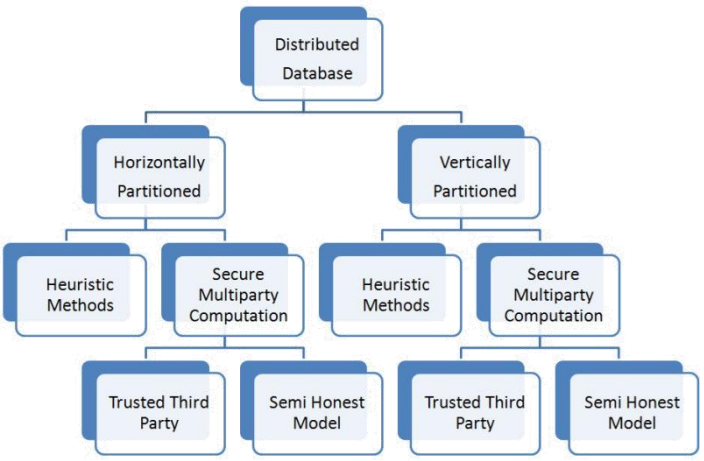


Figure 3: Distributed Database

1. SECURE MULTIPARTY COMPUTATION (SMC) WITH TRUSTED THIRD PARTY

It was Client – Server Architecture bases technique where one party is a client and other parties are clients. All the client parties believe the server party to be trusted and honest such that the server party won't reveal their sensitive and private data to another party [7]. Fig 4 shows SMC with trusted third party.

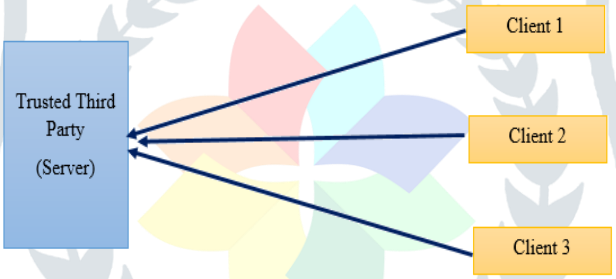


Figure 4: SMC with Trusted Third Party

In this, each party finds the frequent itemset and its local support count and sends it the third party. On receiving these data, the third party evaluates this data to find global frequent itemset and global support count. The result found is returned back to all the client sites for further manipulations [7].

The limitation of this technique was what if the third party fails or collusion between third party and any client [7]. There were more chances of data loss in this technique.

2. SECURE MULTIPARTY COMPUTATION (SMC) WITH SEMI-HONEST MODEL

This technique is quite different than SMC with Trusted Third Party Model. A partially-honest party is one who follows the standard rules. But feels free to migrate in between the steps to gain more information and satisfy an independent agenda of interests [8]. In other words, a partial-honest party follows the rules step by step and computes exactly required values based on the input from the other parties and it can analysis other party's data. But it is sure that it will not insert any false value which results in failure and try to use all the information secured to know sensitive information of other parties. The collusion among the parties does not occur in this model and thus private data is not revealed [8]. Fig 5 shows SMC with the semi honest model.

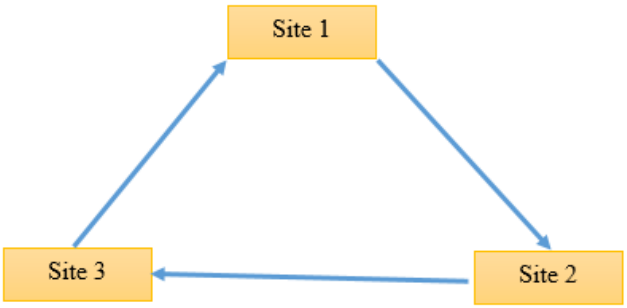


Figure 5: SMC with Semi Honest Model

- Each site assumes the other party to be honest.
- Each site follows the protocol
- Each site computes only the required data.
- Sites do not collude with each other. But try to find some information about other site [13].

The drawback of the model was that each site is assumed to be honest. But there is no surety about sites not colluding with another site.

MAIN TECHNIQUES OF PPDM

Among the two main categories of PPDM approaches viz.

- ✓ **Randomization based**
- ✓ **Cryptography based**

Cryptography based approaches[14], the Secure Multiparty Computation (SMC) provides higher level of privacy but incurs higher computational and communication overhead. As compared, homomorphic encryption based approach provides high level of privacy but incurs higher computational cost. This issue requires critical investigation when applied to data mining. This is so, since data mining requires huge databases as input; hence scalable techniques for privacy pre-serving data mining are needed to handle them. Therefore, in this paper, we mainly focus on reducing the computational cost of privacy preserving data mining algorithm[9].

As discussed, Cryptography based approaches[14] achieve high level of privacy but the resultant protocols are inefficient in terms of computation and communication overhead. As discussed told already, the oblivious transfer based approaches proposed in are not scalable due to their high computational and communicational overhead. Homomorphic encryption based approaches proposed in are computationally expensive due to their complex public key operations. Hence, the scope of above two approaches is limited to small datasets and it is necessary to explore alternative technique that is scalable in terms of dataset size[9].

Solution using our proposed Algorithm:

The secret sharing based approach is an attractive solution for PPDM which greatly reduces the computational and communication cost of SMC and provides high level of privacy. In practical scenario, the assumption about TTP cannot always be ensured and if ensured, compromise in TTP will jeopardize the privacy. Our approach is more relevant in reducing computational cost as compared to communication cost [9].

IV. TECHNICAL PRELIMINARIES

A. Shamir's secret sharing

Shamir's secret sharing method[10] allows a dealer D to distribute a secret value among *n* peers, such that the knowledge of any peers is required to reconstruct the secret. Themethod is described in Algorithm 1.

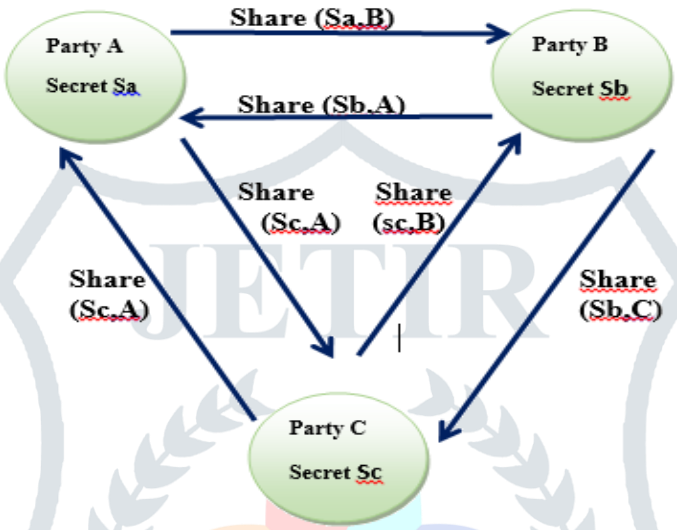
Algorithm 1:(Shamir's secret sharing algorithm)
Require: Secret value *vs*,
P: Set of parties *P*1, *P*2, ・ ・ ・ , *Pn* to distribute the shares,
k : Number of shares required to reconstruct the secret.

1: Select a random polynomial$q(x) = a_{k-1}x^{k-1} + \cdots + a_1x^1 + vs$, where $ak-1 \neq 0$, $q(0) = vs$.
2: Choose *n* publicly known distinct random values

*x*1, *x*2, ・ ・ ・, *xn* such that *xi* ≠ 0.
3: Compute the share of each peer *pi*, where *sharei* = *q*(*xi*).
4: for *i* = 1 to *n* do
5: Send *share i* to peer *pi*.
6: end for.

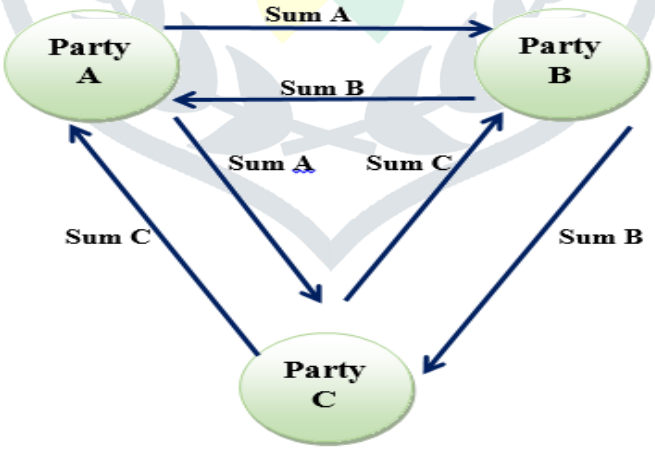
Shamir's method is theoretically secure, in order to construct the secret value *vs* , at least *k* shares are required to determine the random polynomial *q*(*x*) of degree *k*−1 , so the completeknowledge of up to *k*−1 peers does not reveal any information about the secret.

As shown in step 1, each party generates and distributes the shares[11].



Step 1: Share generation And Distribution

In step 2, each party performs the addition of the shares it receives including its ownshare and sends the calculated sum back to every other party. By solving the linear equations corresponding to the received sums, parties are able to calculate the sum of the secret values of all parties using Lagrange's interpolation[11].



Step 2:sending sum of share to each party

B. MFI (Maximal Frequent Item sets)
MFI is a concept used in frequent item sets mining. A frequent item set is called maximal if it is not a subset of any other frequent item set [12]. The set of all maximal frequent item sets is denoted as MFI. An important property is that |MFI|<|FI|.

V. THE PROPOSED ALGORITHM

INIT PHASE:

Initiator sends RSA's public key to all other sites.

FIRST PHASE:

- Each site independently and parallel finds its local MFI.
- Encrypt using RSA public key.
- Find Global MFI
- Send Global MFI to each site.
- Find frequent item count based on global MFI.

SECOND PHASE:

Privacy-Preserving Algorithm to Collaboratively Compute c.count

Require: P : Set of parties $P_1, P_2, \cdots P_n$.
A_{ij} : Secret value of P_i,
X: A set of n publicly known random values $x_1, x_2, \cdots, x_n$.
k: Degree of the random polynomial $k = n - 1$.

1: For each transaction $i = 1$ to N do

2: For each party P_i, $(i = 1, 2, \cdots, n)$ do

3: Select a random polynomial $q_i(x) = a_{n-1}x^{n-1} + \text{-----} + a_1x^1 + A_{ij}$

4: Compute the share of each party P_t, where $share(A_{ij}, P_t) = q_i(x_t)$

5: For $t = 1$ to n do

6: Send $share(A_{ij}, P_t)$ to party P_t

7: Receive the shares $share(A_{ij}, P_t)$ from every party P_t.

8: Compute $S(x_i) = q_1(x_i) + q_2(x_i) + \cdots, +q_n(x_i)$

9: For $t = 1$ to n do

10: Send $S(x_i)$ to party P_t

11: Receive the results $S(x_i)$ from every party P_t.

12: Solve the set of equations to find the sum $= \sum_{i=1}^{n} Aij$ of secret values.

13: If the $\sum_{i=1}^{n} Aij = n$, let $mj = 1$, otherwise $mj = 0$.

14: Each party computes $c.count = \sum_{i=1}^{n} mj$

Example:

Assume that there are 4 parties $P1, P2, P3, P4$ with secret values $A1j = 1, A2j = 0, A3j = 1, A4j = 1$ respectively for arbitrary transaction j in database DB, and that they want to decide whether the transaction j supports the association rule without revealing their values to each other[11].

At first, they decide on a polynomial degree $k = 3$ and $m = 4$ publicly known distinct random values $X = (2, 3, 5, 6)$.

Each party Pi then chooses a random polynomial $qi(x)$ of degree $k = 3$ whose constant term is the secret value $Aij[11]$.

$P1$ picks $q1(x) = x^3 + 3x^2 + 2x + 1$ and computes the shares for other parties such that the share of party Pt, $share(A1j, Pt) = q1(xt)$, where xti s the tth element of X. Thus the shares computed by $P1$ are as follows:

$$Share\ (A1j, P1) = q1(2) = 25,$$
$$Share\ (A1j, P2) = q1(3) = 61$$
$$Share\ (A1j, P3) = q1(5) = 211$$

$$Share\ (A1j, P4) = q1(6) = 337$$

During the second phase, each party adds up all the sharesreceived from other parties and then sends this result to allother parties. That is, party *Pi* computes $S(xi) = q1(xi) + q2(xi) + q3(xi) + q4(xi)$ and sends to all other parties.

At the third computation phase, each party *Pi* will have the4 values of polynomial $S(x) = q1(x)+q2(x)+q3(x)+q4(x) = b^3x_3 + b^2x_2 + b^1x + b$ at $X = (2, 3, 5, 6)$ with the constant term equal to the sum of all secret values. So each party *Pi* can get liner equations:

$$\begin{cases} 8b3 + 4b2 + 2b1 + b = 95 \\ 27b3 + 9b2 + 3b1 + b = 237 \\ 125b3 + 25b2 + 5b1 + b = 863 \\ 216b3 + 36b2 + 6b1 + b = 1407 \end{cases}$$

and get $b = 3$ through solving the above liner equations[11], so

$$\sum_{i=1}^{4} \text{Aij} = \text{b} = 3$$

Because the $= \sum_{i=1}^{n} Aij$=b=3, it means the values of *P*1, *P*2, *P*3, *P*4 in transaction *I* of database *DB* are not all 1,so the transaction *i* does not support the association rule, then let *mj* = 0.

V. THEORETICAL ANALYSIS

Several metrics for evaluating privacy preserving data mining techniques are dis-cussed in [13]. Based on this, we analyze our approach for privacy, correctness, computation cost and communication cost.

Privacy

In our proposed approach, the secret value FI of a party Pi cannot be revealed even if all the remaining parties exchange their shares. Since each party Pi executes Shamir's secret sharing algorithm with a random polynomial of degree n-1, the value of that polynomial at n different points are needed in order to compute the coefficients of the corresponding polynomial, i.e., the secret value of party Pi. Pi computes the value of its polynomial at n points as shares, and then keeps one of these shares for itself and sends the remaining n-1 shares to other parties. Since all n shares are needed to reveal the secret, other parties cannot compute secret even if they combine their shares.

Further, no party learns anything more than its prescribed output. This is so, be-cause as per the approach followed every party shares its FI as the secret; for which it chooses different polynomial randomly. Hence, it is not possible for a party to determine the secret values of other parties, since the individual polynomial coefficient selected by each party is not known to other parties[9].

Computation Cost (Time Factor)

The computation cost depends on the FI and the no. of iterations required for finding final FI. We give here the computation cost for single iteration. As-sume that for every party Pi, the cost of generating random polynomial qi(x), i = 1, 2,..., n is C. In proposed approach, we have two values as a secret so we have to gen-erate random polynomial two times.. The total number of $2n\ (n - 1)$ additions are calculated to find s(x) = q1(x) + q2(x) +,…,+qn(x). Efficient O(nlog2n) algorithms for polynomial evaluation are available [9]. Hence the compu-tation cost for our proposed approach is quadratic i.e. O(n2).

Communication Cost (Number Of message Transfer)

In comparison to Trusted Third Party based approach, our approach incurs more communication cost because for collaboratively computing Frequent item, communication between every party is necessary.

VI. CONCLUSION

From our work, we can conclude that by applying the Shamir's secret sharing on distributed Association mining algorithm, we remove need of trusted third party in distributed database.

We compared our approach with the oblivious polynomial evaluation and holomorphic encryptions based approaches proposed in and show that in terms of computation cost, our approach is hundreds of magnitude faster than the oblivious polynomial evaluation and holomorphic encryption based approaches and hence is more suitable for large datasets in practical scenario [9][11].

The performance is optimal on all the datasets, even if the number of sites are increased.

ACKNOWLEDGEMENT

We are deeply indebted & would like to express gratitude to our thesis guide Prof. Nikul Virpariya , B. H. Gardi College of Engineering & Technology for his great efforts and instructive comments in the dissertation work.

We would also like to extend our gratitude to Prof.Hemal Rajyaguru, Head of the Computer Science & Engineering Department, B. H. Gardi College of Engineering & Technology for his continuous encouragement and motivation.

We would also like to extend our gratitude to Prof. Vaseem Ghada, PG Coordinator, B. H. Gardi College of Engineering & Technology for his continuous support and cooperation.

We should express our thanks to our dear friends & our classmates for their help in this research; for their company during the research, for their help in developing the simulation environment.

We would like to express our special thanks to our family for their endless love and support throughout our life. Without them, life would not be that easy and beautiful.

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] Shariq J. Rizvi And Jayant R. Haritsa, Maintaing"*Data Privacy In Association Rule Mining," In Proceedings Of The 28th International Conference On Very Large Databases 2002.*

[6] Duraiswamy, K., Manjula, D.: *Advanced Approach In Sensitive Rule Hiding. Modern Applied Science 3(2) 2009*

[7] N V Muthu Lakshmi and Dr. K Sandhya Rani," *Privacy Preserving Association Rule Mining Without Trusted Party For Horizontally Partitioned Databases"International Journal of Data Mining AND Knowledge Management Process (IJDKP) Vol.2, No.2 March 2012*

[8] J. Han and M. Kamber, Data Mining: *Concepts and Techniques. San Fran-cisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 227-245.*

[9] Sankita Patel, Sweta Garasia And Devesh Jinwala" *Preserving Privacy K-Means Clustering Based On Shamir's Secret Sharing Scheme" adfa, p. 1, 2011. © Springer-Verlag Berlin Heidelberg 2011*

[10] Xinjing Ge, Li Yan†, Jianming Zhu‡, Wenjie Shi"*PPARM Based On The Secret Sharing Technique"IEEE,*

[11] Sankita Patel, Viren Patel, and Devesh Jinwala," *Privacy Preserving Distributed K-Means Clustering in Malicious Model Using Zero Knowledge Proof" C. Hota and P.K. Srimani (Eds.): ICDCIT 2013, LNCS 7753, pp. 420–431, 2013.© Springer-Verlag Berlin Heidelberg 2013*

[12] Murat Kantarcioglou And Chris Clifton,,"*Privacy-Preserving Distributed Min-Ing Of Association Rules On Horizontally Partitioned Data", In Proceedings Of The Acm Sigmodworkshop On Research Isuues In Data Mining And Knowledge Discovery (2002), 24-31.*

[13] Liwu Chang And Ira S. Moskowitz,"*Parsimonious Downgrading And Decision Trees Applied To The Inference Problem,"In Proceedings Of The 1998 New Security Paradigms Workshop, 82-89. 1998*

[14] William Stallings, "*Cryptography and Network Security", Fifth Edition, 2011*

[15] J. Han, M. Kamber, In "*Data Mining: Concepts And Techniques", 2nd Edition;Morgan Kaufmann Publishers Is Imprint From Elsevier, San Francisco, Ca 941*

AUTHOR PROFILES:

1)Manvar Sagar is a student of Masters of Engineering in Computer Engineering at B. H. Gardi College of Enggineering and Technology, Rajkot, Gujarat, India. He is bachelors in Information Technology. His area of interest are Data Mining,Computer networking and Security.
Contact:+91 9586507454

2)**Desai Sonali** is a student of Masters of Engineering in Computer Science and Engineering at B.H.Gardi College of Engineering and Technology, Rajkot, Gujarat, India. She is bachelors in Computer Science and Engineering. Her area of interest are Data Mining, Computer networking and Security.Contact:+91 9408966536

3)Hathi Karishma is a student of Masters of Engineering in Computer Science and Engineering at B.H.Gardi College of Engineering and Technology, Rajkot, Gujarat, India. She is bachelors in Computer Science and Engineering. Her area of interest are Data Mining, Computer networking and Security. Contact:+91 9429810304