Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Augmented Privacy of Association Rule Mining among Horizontal and Vertical Partitioned Databases

[1]Mule. Taruni, [2]P. Jeevana Jyothi
[1](M.Tech) - CSE, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India
[2]Assistant Professor, Dept of CSE, Vasireddy Venkatadri Institute of Technology (VVIT), Andhra Pradesh, India

Abstract: ***Data mining is most rapid mounting area today which is used to extract important knowledge from large data collections but often these collections are divided among several parties. Privacy liability may prevent the parties from directly sharing the data and some types of information about the data. Internet today has put up a great challenge on the security for Indian bank Sector. In today's growing environment, most of the computation is jointly (Union) computed involving inputs of all the banks. Such computations use confidential data of the involved banks to compute the result. Each bank is having confidential data which they would not like to share with other banks. Privacy preservation is of great concern as no banks can be trusted in real scenario. In this paper we have proposed an efficient innovative secure protocol for computation. This paper is an extension of our presented work in which leakage of information may be possible due to insufficient security, simplicity and efficiency are not well in the databases. For the sake of secure computation of associated private subsets that are held by the other parties. This paper augment the association rule mining in both horizontal and vertically distributed databases and contemplates the in depth ways used for mining association rules over distributed where as to maintain privacy and this paper uses A Secure Trusted multiparty Protocol which performs associated bankers data securely and efficiently.***

Keywords: ***Secure Multiparty Computation (SMC), Secure Trusted multiparty Protocol (STMPS), Single STMPS, multi STMPS, privacy, security***

I. INTRODUCTION

Data mining can extract important knowledge from large data collections but sometimes these collections are split among various parties. Privacy liability may prevent the parties from directly sharing the data, and some types of about the data. Data mining technology has become prominent as a means of identifying patterns and trends from large quantities of data. In secure association rule mining for databases, given a set of transactions, where each transaction is a set of items, an association rule is an expression X->Y, where X and Y are sets of items. The intuitive meaning of such a rule is that the transactions that contain the items in X tend to also contain the items in Y. Data mining and data warehousing co-jointly: most popular tools operate by gathering all data into a central database (Master Database) then running an STMP algorithm against that data. However, privacy liability can prevent building a centralized warehouse, data may be distributed among several custodians none of which are allowed to transfer their data to another site. In Joint computation is a need in fast growing Internet world. Most of the applications work on joint computation where large numbers of parties are involved. These parties send their data for computation to STMPs and the computing STMPs announces the result. The first major concern here is to maintain privacy of inputs provided by the parties. Security and correctness in the result of computation is the next parameter which has to be maintained in the protocol. This problem is SMC, where n parties (BANKS) send their private inputs x1, x2,… xn to STMPs for computation and STMPs announces the result in form of y. The general Secure Multiparty Computation (SMC) model for Indian BANK sector is:

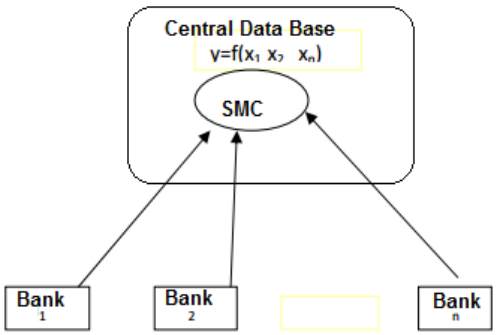

Figure1. Basic SMC Model

In figure 1 n BANKS send their inputs (Secured Customer data) to Central Database for computation. The major concern here is to maintain privacy, security and correctness in the result. Privacy preserving data mining solutions are of great importance in the field of banking operations. Consider a case where numbers of different banks wish to jointly mine

their Customer's data for the sake of generating loan avail eligibility list in this connection when apply the mining technique on master database along with to generate association rules on distributed data and prevention of data is to maintained due to confidentiality of customers record. This problem is an instance of SMC problem. Privacy preserving data mining solutions enable the banks to compute desired data mining algorithms on the union of their databases, without ever disclosing the data. The only information learned by different banks is the output of data mining algorithm.

II. BACKGROUND AND RELATED WORK:

There is an absence of proper data security and cyber laws which is hindering banking's and its business prospects. There is also remarkable excitement and a lack of understanding of the problems surrounding security. The most significant security issue is the protection of data. Some of the information security and data privacy challenges that Banking face include lack of tough data protection laws, use of portable devices such as laptops by employees to store confidential information, rising data security costs due to increased employee background checks, training employees in maintaining data security, ensuring compliance with security policies implemented in the Banking, and problems through employee activity monitoring procedures. To maintain the confidentiality of Customers or Banking information, there is need to implement data security and privacy procedures.

In Secure multiparty computation it is the issue of n gatherings to register a private capacity of their inputs in a safe system, where security implies the right result processed by the STMPs for keeping up the parties' protection as a parties' portion may need to abuse the other party's data. We accept that the inputs are x1,x2,… xn where xi is the data of party Pi and the STMPs will process a capacity f(x1,x2,… .xn)=y and declare the outcome y [1]. Security is intended to accomplish rightness of the consequence of computation and keeping the party's data private regardless of the possibility that a parties' percentage are undermined. In figure 3, trusted outsider is utilized for doing the computation on the inputs gave by the gatherings. As indicated by [2], the real issue with this methodology is it is hard to locate the outsider which is trusted by every one of the gatherings giving the inputs and to control the capacity of enemies. The idea of SMC started in 1982 when Yao proposed and offered answer for his mogul's issue in which two tycoons needed to know who was wealthier without unveiling individual riches to one another [3]. It was a two-party computation protocol for semi legit parties who take after the protocol additionally attempt to know an option that is other than the outcome. The thought was reached out to multiparty computation by numerous researchers.

Goldreich et al. showed the existence of a secure solution of SMC problem. The protocol's span relies on upon the quantity of gatherings included in the computation process. [4].They utilized circuit assessment protocols for secure computation. Prior exploration concentrated on hypothetical studies. Later, some genuine applications developed like Private Data Recovery (PIR) [5, 6], Protection preserving data mining, Security preserving geometric computation [9], Protection preserving experimental computation, Security preserving factual investigation and so on. A point by point survey of Secure multiparty computation exploration is given by Du et al. in which they added to a structure for issue revelation and changing over typical issue to secure multiparty computation issue.

A review of secure multiparty computation with special focus on telecommunication systems is given by Oleshchuk et al. in. Yao's unique protocol considered just the instance of semi-legit parties; an expansion to the instance of malignant party was given by Lindell. Ronald Cramer gave the hypothetical examination of multifaceted nature limitations on secure multiparty protocols, particularly for the mystery sharing issue. Ran Canetti distinguishes blemishes in past multi-party computation work emerging from the presentation of versatile enemies, who decide to degenerate included gatherings progressively amid the computation. The paper presented the thought of a semi-genuine party, who gives off an impression of being straightforward from an outside point of view, however veers off from the protocol somehow. He introduced a protected protocol, utilizing a trusted outsider, to keep away from the versatile antagonistic pitfalls. Going for security preserving registering of factual dispersion, which is much of the time experienced in measurements, and in view of the immovability of processing discrete logarithm and utilizing thorough rationale, they proposed the arrangement. Exhibited the protocols permitting the players to safely take care of standard computational issues in direct variable based math, for example, determinant of grids item, rank of a framework, and focus closeness between matrices.

III. PARTITIONING OF DATABASE

3.1 Database Partition

Association rule mining is one of the Data Mining techniques used in distributed database. In distributed database the data may be partitioned into fragments and each fragment is assigned to one site. The issue of privacy arises when the data is distributed among multiple sites and no other party wishes to provide their private data to their sites but their main goal is to know the global result obtained by the mining process. However privacy preserving data mining came into the picture. As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment.

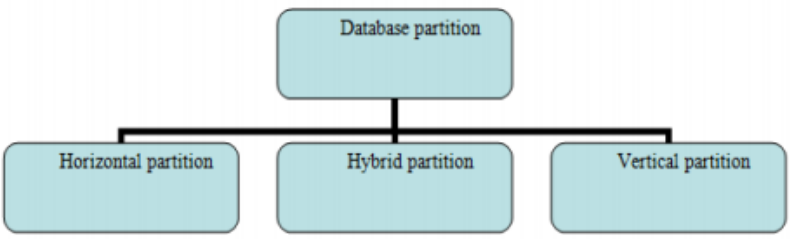

Figure2. Different types of partition of Database

Here in this section database will be partitioned as horizontal and vertical view, Customers details $C_{i....}C_n$ Will be stored in the database with various field attributes like Customer name ,Customer account number ,gender, age ,profession ,available balance, loans with bank, account status etc...

Table1. Master Database of All customers from different bank

Vertical Distribution

Horizantal Distribution

S.No	Customer account number	Customer name	gender	age	profession	available balance	loans with bank	account status
1	1211	a	M	20	Student	1000	---	Active
2	1213	b	F	60	Professor	100000	Axis	Active
3	1214	c	M	40	Doctor	1000000	SBI	Active
4	1215	d	F	25	lawyer	100000	---	Active
5	1216	e	M	39	Farmer	10000	___	Active

From the above table we need to partition database as vertical view in this connection vertical database selection done using association rules based on parameters like customer who are having male and age>20 by applying where condition on master database we will get details as table 2.

3.2 Vertical Partitioning:

Table 2. Vertical Data distribution using association rules

S.No	Customer name	Gender	age
1	c	M	40
2	e	M	39

As showing above horizontal database will partition where to display the customer account details from account number from 1211-1214. Then it shows all the fields of customers data in horizontal records well be displayed as shown in table 3.

3.3 Horizontal Partitioning:

Table3.Horizontal Data Distribution using association rules

S.No	Customer account number	Customer name	gender	age	profession	available balance	loans with bank	account status
1	1211	a	M	20	Student	1000	---	Active
2	1213	b	F	60	Professor	100000	Axis	Active
3	1214	c	M	40	Doctor	1000000	SBI	Active

Above database partition will be done based on applying association rules on data mining.

IV. PROPOSED SYSTEM

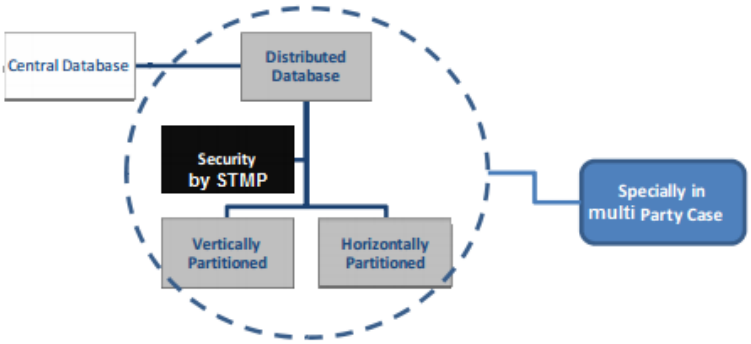

Figure3. STMP Protocol on Partitioned database

In this protocol all the banks involved in computation split their data into x data fields and encrypt data through some pre-decided encryption method. The encrypted data Eij is send to third party (who generates reports from master database). This is an untrusted layer whose task is to forward the data fields to STMPs selected at runtime for computation. Third party cannot store the data, they just forward it. As report generators are untrusted, so they hold data fields of the customers and not the entire data. After computation majority of STMPs giving the same result is considered as the right result of computation as correctness is a major parameter for computation which has been analyzed in previous work.

4.1 Three layer architectural framework:

1. n Banks : B1, B2… Bn with user data fields xij
2. Report Generator (Third party layer)
3. Multiple STMP layer: STMP1, STMP 2… STMP n

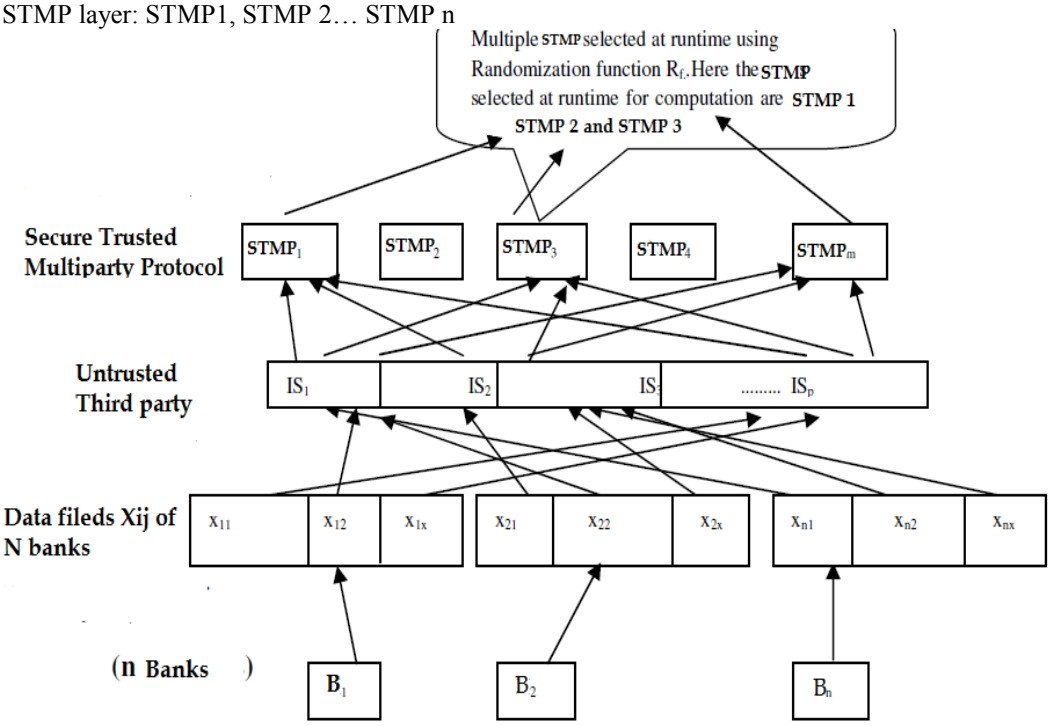

Figure 4. Three layer architectural SMC framework for banks using Secure Trusted multi party computation

4.2 Formal description of algorithm

- Bi – Banks where i ranges from 1 to n
- xij – Data of customers Hi where j ranges from 1 to x
- Rij – Random data of customers Hi where j ranges from 1 to q
- Dij – total data including the random and the original data
- Eij – Encrypted data associated with Customers Hi where j ranges from 1 to x+q
- ISp– untrusted third party, where p ranges from 1 to z
- STMP – Secured Trusted Multiparty

4.3 STMP Algorithm:

- Generate xij fields for every bank Hi
- Generate random data Rij for every xij
- Group random data Rij with original data xij to get Dij
- Encrypt data Dij using pre-decided encryption method to get Eij.
- Distribute the encrypted data Eij among the Third party Ap
- Send the data from un-trusted Third party Ap to STMPs
- Calculate the result at STMPs using the encrypted data and the keys.
- Identify the STMPs at runtime for performing computation.
- The result is announced by STMPs
- Majority of STMPs giving same identical result is considered as correct result.

V. SECURITY ANALYSIS

Privacy preserving data mining is defined as preserving the individual privacy and retaining the information in dataset to be released for mining. If the Trusted third party is malicious then it can reveal the identity of the source of data. A set of third party from the inscrutable layer will make the source of data ambiguous and will preserve the privacy of individual. The more the number of third party in the inscrutable layer the less will be the possibility of hacking the privacy of the

data. The third party hide the identity of the bank. In the protocol there is one layer of third party, consisting of p third party IS1, IS2, IS3…, ISp. Then the probability of revealing the source of the data at TTP is inversely proportional to the number of parties sending data. We can see that there is more security when there are large numbers of participants.
The probability of hacking the data of a single bank Hi:
P_ Bi_ = 1/n …… (1)
Where n is total number of banks involved in computation.
The probability of hacking data of r banks:
P (Br) = r/n…… (2)
Therefore, total Probability for leak of the data fields
= [r/n] * [r=1_r X r)/ (r=1_n X r] …….. (3)
Where X r are the data fields of r banks.

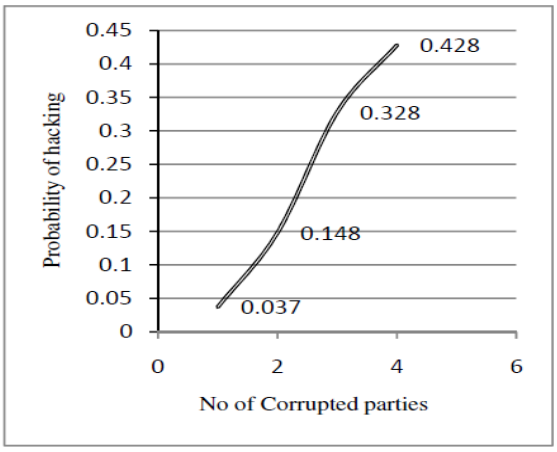

Figure 5. Security analysis with increased number of corrupted parties

Drawbacks with Existing Protocol (SMC)
With our presented protocol we have noticed some of the drawbacks

- leakage of information may be possible due to insufficient security, simplicity and efficiency are not well in the databases

Advantages of Proposed Protocol (STMP)
Use of Secured Trusted Multiparty Protocol (STMPs for computation makes correctness parameter to be more truthful.

- Providing encrypted inputs to Third party makes the protocol secured.
- Use of Third party layer makes the identity of Banks ambiguous as privacy of inputs is a major concern.
- Splitting of inputs into data fields and random distribution of it to Third party increases The security parameter.
- Probability of malicious conduct increases exponentially with increased number of Corrupted parties.
- The attractiveness of this protocol is: STMPs are performing the same Computation and announce the result. With the help of such computation, malicious STMP can be easily traced after several round of computation.

Using this protocol and algorithm a wide variety of computations can be optimally performed with enhanced security and privacy.

VI. CONCLUSION

In this paper we have proposed an efficient innovative secure protocol for computation it is an extension of our presented work in which leakage of information may be possible due to insufficient security, simplicity and efficiency are not well in the databases. For the sake of secure computation of associated private subsets that are held by the other parties. This paper uses A Secure Trusted multiparty Protocol which performs associated bankers data securely and efficiently.

REFERENCES

[1] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014
[2] C. Clifton, M. Kantarcioglu , J. Vaidya, X. Lin and Y. Michael. (2002), Tools for privacy preserving distributed data mining , SIGKDD Explorations Volume – 4,Issue – 2, 1-8.
[3] J. Vaidya, and Chris Clifton. (2003), Leveraging the Multi in Secure Multi-Party Computation, in the proceeding of the 2003 ACM workshop on privacy in electronic society, ACM Press.
[4] A.C.Yao. (1982), Protocol for secure computations, in Proc. 23rd IEEE Symposium on the Foundation of Computer Science (FOCS), IEEE, 160-164.

[5] O. Goldreich, S. Micali, and A. Wigderson. (1987), How to play any mental game, in STOC '87: Proceedings of the nineteenth annual ACM conference on Theory of computing, New York, NY, USA: ACM, 218-229.

[6] B.Chor and N.Gilbao. (1997), Computationally Private Information Retrieval (Extended Abstract), in proceedings of 29th annual ACM Symposium on Theory of Computing, El Paso, TX USA.

[7] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. (1995), Private Information Retrieval, in proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science, Milwaukee WI, 41-50.

[8] Y. Lindell and B. Pinkas. (2000), Privacy preserving data mining, in advances in cryptography- Crypto2000, lecture notes in computer science, vol. 1880,2000.

[9] R. Agrawal and R. Srikant. (2000), Privacy-Preserving Data Mining, in proceedings of the 2000 ACM SIGMOD on management of data, Dallas, TX USA, 439-450.

[10] M. J. Atallah and W. Du. (2001), Secure Multiparty Computational Geometry, in proceedings of Seventh International Workshop on Algorithms and Data Structures(WADS2001), Providence, Rhode Island, USA,165-179.

[11] W. Du and M.J. Atallah. (2001), Privacy-Preserving Cooperative Scientific Computations, in 14th IEEE Computer Security Foundations Workshop, Nova Scotia, Canada, pages 273-282, Jun. 11-13 2001.

[12] Assaf Schuster, Ran Wolff, Bobi Gilburd," Privacy-Preserving Association Rule Mining in LargeScale Distributed Systems", fourth IEEE symposium on Cluster Computing and Grid, 2004.

[13] Tirumala prasad B, Dr. MHM Krishna Prasad, "Distributed Count Association Rule Mining Algorithm", International Journal of Computer Trends and Technology, July to Aug Issue 2011, pp.280-284.

[14] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. "Association Rule Hiding for Data Mining", Springer Series: Advances in Database Systems, Vol. 41, 1st Edition., 2010, p.13.

