ISSN: 2321-6840

Privacy-Preservation in Collaborative Association Rule Mining for Outsourced Data

Khushbu Agrawal[1] and Vandan Tewari[2]
[1]M.E. Scholar, C.S.E. Department, S.G.S.I.T.S., Indore, Madhya Pradesh, India.
Email: khush.agrawal08@gmail.com
[2]Assistant Professor, C.T.A. Department S.G.S.I.T.S., Indore, Madhya Pradesh, India.
Email: vandantewari@gmail.com

Abstract: In recent years, the explosion of digital data and in-formation, and various applications such as real-time monitoring, distributed collaboration, large-scale medical and financial data analysis and social network accumulate large amounts of data from different data owners. The burgeoning ability to generate vast volumes of data presents technical challenges for efficient data mining. Meanwhile, with the emergence of cloud computing and its model for IT services, which affords both computational and storage scalability, the outsourcing of data for storage and mining services is acquiring popularity. So, many organizations having insufficient storage and computational resources, willing to reduce their storage and computation cost, are widely adopting the outsourcing of the data mining jobs to a third party service provider. These service providers are assumed to be semi-trusted parties for privacy concerns. In this paper, we propose a collaborative privacy-preserving data mining (CPPDM) solution for outsourced data, which ensures that the data is stored, processed and shared without violating the user privacy. In our solution, we are using anonymization and encryption techniques for user privacy.

Keywords: CPPDM, Outsourcing, Privacy-preservation, Semi-trusted, Encryption, Anonymization, MapReduce.

I. Introduction

Frequent itemset mining and association rule mining, two popular, well-designed and widely adopted data analysis methods, are generally used for getting frequently co-occurring data items and significant relationships between seemingly unrelated data respectively in a large transaction databases. These two methodologies have been employed in a range of applications, such as market basket analysis, prediction, health care, web-usage mining, bioinformatics and so on [23].

Let I = {i1, i2,…., in}, be a set of data items and D = {t1, t2, ….., tm}, be a set of transactions called the database. In database D, each transaction has a unique transaction ID and contains a subset of items in I. In order to obtain frequent itemsets from set of all data items, some constraints are im-plied. The best-known constraints are minimum threshold on support and confidence. Support indicates that how frequently the itemset appears in the database. An itemset is said to be frequent if and only if the support of the itemset is greater than or equal to minimum support (Ts). An association rule is represented as X => Y, where X, Y $\subset$ I. An example rule for the supermarket could be {bread, butter} => {milk} meaning that if bread and butter are bought, customers may also buy milk. X => Y is significant and useful if the confidence is high (≥Tc). Confidence indicates how often the rule has been found to be true. After mining frequent itemsets and their supports, it is easy to mine association rules. However, such old but familiar and important techniques are facing new challenges and problems nowadays.

In recent years, huge amount of data is collected from different data owners through various applications such as real-time monitoring, distributed collaboration [6], large-scale medical and financial data analysis [3], and social network [22], [2], [4], [5]. Meanwhile, cloud computing came into existence and started changing people's everyday life. With its computation and storage scalability, cloud computing attracted more and more data owners to reduce their expensive storage and management cost. Data owners choose to use data mining as a service by outsourcing their rule mining needs to third party service provider. Some of the popular cloud service providers are Amazon S3, Microsoft Azure, and Google App Engine. They provide cost-effective and user-friendly cloud services.

Although these service providers have the vested interest in providing the security and privacy to user data, there are instances where user data are shared or compromised. Hence, data owners cannot trust online service providers blindly, but assume them as semi-trusted parties. Storing sensitive data in an semi-trusted storage server leads to privacy violations. Privacy and security issues are inevitable and the primary concern of the Data-Mining-as-a-service paradigm. There are scenarios where the data owner does not want to share their private and sensitive data either with third party service provider or with other data owners.

In this paper, we propose a collaborative privacy-preserving frequent itemset mining solution for outsourced data, which

is then used to prepare a privacy-preserving association rule mining solution. In our solution for user privacy, we are using anonymization and encryption techniques. Data owners can outsource their encrypted data and mining jobs to a semi-trusted third party in a privacy-preserving manner. Both the solutions are designed for a range of applications where a high level of user privacy is required.

In the following sections, we will present and analyze a set of privacy preserving mechanisms and some well-known applications of outsourced data mining. In Section II, motivation for our study is presented. In Sections III, we present the models and design goals for our study. Previous work related to our study is discussed in Section IV. In Section V, we present our solution for privacy preservation in outsourcing data mining scenario. The required background content is also discussed in this Section. Section VI presents the privacy analysis regarding our proposed approach. The experimentation details and results are presented in Section VII. The conclusion of the paper is presented in Section VIII.

II. Motivation

In recent years, as outsourcing of data mining jobs to third parties has become very popular, privacy and security issues have also become the primary concern for data owners. This has forced the researchers and industry to continuously look for the new secure and privacy-preserving outsourced data mining solutions. In this paper, our objective is to further explore what, why, and how of privacy preservation in outsourcing data mining scenario. Previously proposed solutions were suitable only for distributed environment, but not for outsourcing scenario. In this paper, we propose a solution that preserves privacy among multiple data owners that is suitable for distributed as well as outsourcing scenario and also makes data owners ensure of minimum leakage of raw data.

III. Models And Design Goal

In this section, we formalize the system model and privacy model used in our study, and identify the design goals.

A. System Model

The system model consists of a server and two or more data owners. Each data owner wants to reduce the cost of computing and storing large amount of their private data, by outsourcing that data to the third party server. In this paper, these server are assumed as semi-trusted. So, privacy becomes inevitable issue regarding outsourcing. Therefore data owners apply anonymization and encryption to their database prior to outsourcing the data to the server. Data owners can also request the server to mine frequent itemsets or association rules from joint encrypted database. The server has the task of storing and compiling the databases received from multiple data owners, the mining of frequent itemsets or association rules from joint encrypted database and sending the mining results to the pertinent data owners.

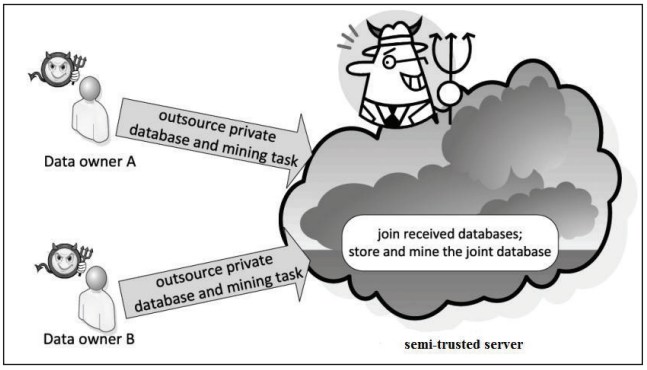

Fig. 1: System Model of Outsourced Data Mining

B. Privacy Model

The third party server is considered as semi-trusted in this paper. Some services provided by them are free of cost, while some are chargeable. There are some instances, where the user data from the server got leaked or hacked or compromised. Hence user-server privacy is considered in our algorithm.

Multiple data owners participate in the collaborative outsourced data mining in order to obtain mining results. In this paper, we assume that each data owner has little information about other's private data, and may be interested to gain more for their financial profits. Hence user-user privacy is also considered in our algorithm.

C. Design Goals

The design goals for our proposed algorithm are as follows:

1) *Privacy:* In today's digital world, data is most valuable thing. Every data owner wants to keep his sensitive data safe from the other data owners as well as the server. So, our design requirements are user-server privacy and user-user privacy.
2) *Accuracy:* Usually, privacy-preservation in outsourced data mining results in less accuracy, and therefore, any trade-off has to be realistic.

IV. Related Work

Privacy preservation in data mining has been a very interesting topic for study for last two decades, and a number of solutions have been proposed [7], [8], [9]. In recent years, privacy preservation in outsourcing data mining scenarios has introduced a new way for study in [10] and [11]. Previously proposed solutions can be broadly categorized into two categories: data perturbation

and secure multi-party computation. In the first method, raw data is transformed by adding some random noise to it, so that it no longer exhibits sensitive information, while the statistical characteristics of the data are preserved. These methods may result in unexpected impacts on data mining preciseness. While, the second method can be used in distributed environment only,

where multiple parties agree to share their part of private data and cooperate to get the required results. However, these methods are not suitable for outsourcing scenarios.

As mentioned above, numerous privacy-preserving association rule mining approaches have been proposed [12], [13], [14], [15], [16], [17]. In [12], one data owner works as the master while other work as slaves. All slaves insert some fake data to their private datasets and send to the master. Every slave also sends his collection of real transaction's IDs to the semi-trusted master. The master mines the association rules from the joint database containing fake data. Like our solution, for mining purpose, a semi-trusted third party is used, but unlike our solution, most of the computational works are performed by a single data owner who is assigned as master. Hence, this solution is hardly suitable for outsourcing mining scenario. Though some fake data (i.e. noise) is inserted in datasets to minimize the usability of data, the master is still able to learn sensitive information from received datasets. It may also affect the data mining precision. In contrast, in our solution, we are not adding noise to the database, but using anonymization and encryption schemes. Our solution does not depend on one particular data owner to perform computations; hence it lowers the possibility of leakage of information.

In [13] and [14], asymmetric homomorphic encryption is used to calculate the supports of item sets, while in other solutions [15], [16], [17], a secure scalar product protocol or a secret sharing scheme is used for these calculations. In These solutions, exact supports of item sets are disclosed to all the data owners, so the information about raw data is got leaked. The Solution presented in [14] does not reveal exact supports, so computation of confidences will be very complicated. However, with this solution association rules can-not be mined. Unlike existing solution based on asymmetric homomorphic encryption, our solution uses RC4 symmetric encryption scheme for secure exchange of keys.

Existing solutions based on privacy-preserving association rule mining [18], [19], [20], [21] have been presented in the setting of the single data owner. Usually, the data owners encrypted the data items with the substitution cipher prior to outsourcing. A substitution cipher is vulnerable to frequency analysis attack, a solution is proposed in [18]. However, later this solution has been proved as not secure in [19]. In [20] and [21], Giannotti et al. proposed a solution based on k-anonymity. Each data owner adds fake transactions so that every item shares the similar frequency with at least k-1 other items. They send encrypted databases containing both the real and fake transactions to the server. The server mines the joint dataset. This solution may affect the mining precision.

V. Proposed Solution

In this section, we present our solution for privacy preservation in outsourced data mining scenario. We present our system architecture followed by detailed description of the solution.

A. System Architecture

Our solution for collaborative association rule mining on outsourced data involves three modules, namely: preprocessing, outsourcing and mining.

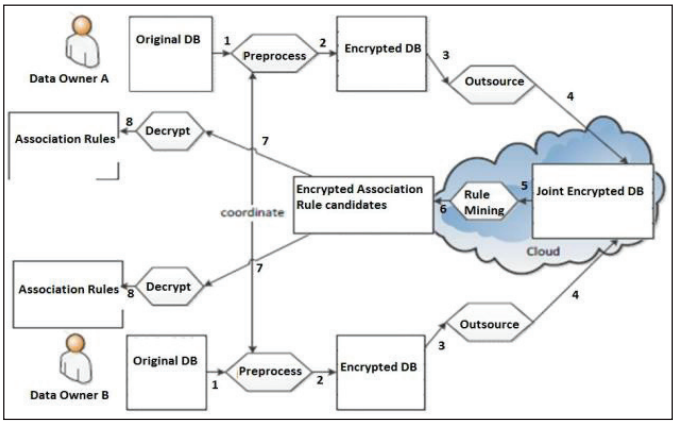

Fig. 2: Proposed System Architecture of Privacy-Preserving Collaborative Association Rule Mining

The preprocessing module is the first step which makes the data ready to be shared with other users of the system and is performed at the client side. Initially, each data owner has its own private transaction database containing customer IDs and item IDs purchased by each customer. In this step, each data owner anonymizes each item by applying the MD5 algorithm and maintains a record of each item with its corresponding message digest value. All customer IDs are encrypted with the RC4 symmetric encryption algorithm. After applying anonymization and encryption on original database, each data owner outsources his preprocessed dataset to the third party (server).

Table I: Original Database (Before Preprocessing)

Customer ID	Item ID
10001	bread, butter
10002	milk, bread
10003	milk
10004	milk, butter, bread
10002	beer
10003	bread, butter, beer
10001	milk, butter
10004	milk

Table II: Preprocessed Database For Outsourcing

Customer ID	Item ID
Mg9BBRnC	d131dd02c5e6eec4, 693d9a0698af95c
MqNBDhmA	ae6dacd436c919c6, d131dd02c5e6eec4
NwxNJAnQ	ae6dacd436c919c6
MA5BDhZC	ae6dacd436c919c6, 693d9a0698af95c, d131dd02c5e6eec4
MqNBDhmA	4004583eb8fb7f8
NwxNJAnQ	d131dd02c5e6eec4, 693d9a0698af95c, 4004583eb8fb7f8
Mg9BBRnC	ae6dacd436c919c6, 693d9a0698af95c
MA5BDhZC	ae6dacd436c919c6

In outsourcing module, data owners having insufficient computational resources and lack of data mining expertise, choose to outsource their data storage, management and mining demands to the third party server. After preprocessing the dataset, each data owner uploads the preprocessed databases to third party server.

In this module, all the computation is performed at the server side. All the received individual preprocessed databases are combined into a joint database. At server side, the joint database is processed through a distributed file system. Hadoop provides a distributed file system, i.e. HDFS and a frame-work for analyzing and transforming large volumes of data sets using the MapReduce paradigm. Hadoop runs classical Apriori algorithm in a parallel manner using the MapReduce framework, so the rules can be generated much faster and in an efficient way. The complete procedure is described here.

The transaction dataset provided as input is divided into multiple input chunks of by default size of 64MB. Each mapper reads all the transactions from its individual input chunk sequentially. Then each mapper computes the frequency of items to create local 1-itemsets. The output of different mappers are combined and sorted. This sorted output is given to the reducer and reducer generates the count of the occurrence of the global 1-itemsets. Then these counts are compared with the threshold value. The itemsets with the count greater than or equal to the threshold are kept and are known as frequent 1-itemsets. All infrequent items are pruned from the database. After pruning, the generated itemset are called k-itemset, where k represents the number of frequent items present in the itemset.

For each itemset, all the possible combinations of the items present in the itemset are generated. The count of these combinations is checked against the provided threshold value. If count satisfies then it is kept in the list of frequent itemsets. This process repeats for each itemset present in the k-itemset. Hence all the frequent itemsets present in the input dataset are generated. The MapReduce flow of the proposed framework is shown below:

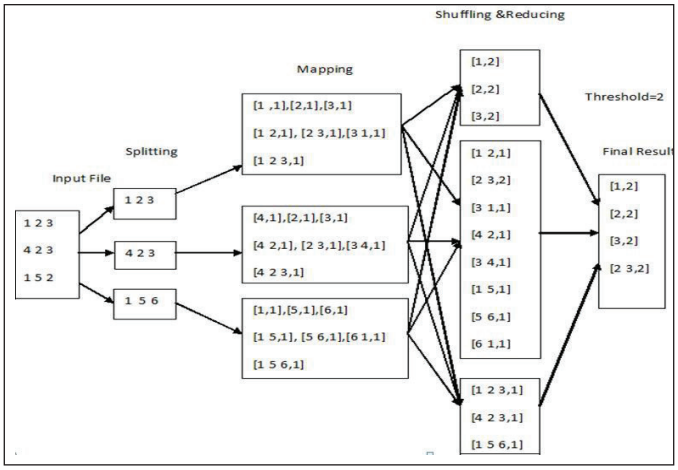

Fig. 3: Apriori Algorithm in MapReduce Framework

The cloud sends the encrypted rules to the corresponding data owners. Data owners decrypt the encrypted rules and finally get the collaboratively mined association rules in privacy-preserving manner.

B. Background

1) *RC4 Encryption:* The RC4 Encryption Algorithm is a shared key stream cipher algorithm that requires interchanging the shared key in a secure manner. It is a symmetric key algorithm, in which data stream is simply XORed with the generated key sequence. The RC4 encryption algorithm is used by standards such as IEEE 802.11 within Wireless Encryption Protocol using 40 and 128-bit keys. It is also used in many commercial software packages such as Lotus Notes and Oracle Secure SQL. RC4 Encryption is about 10 times faster than DES.
2) *MD5 Algorithm:* The MD5 algorithm is a widely used hashing algorithm. It is a one-way cryptographic function that accepts a message of arbitrary length as input and returns a fixed-length (128 bit) digest value as output to be used for authentication of the original message. To be considered cryptographically secure, produced digests should be random. No message can be generated, matching a specific hash value; and it is impossible to create two messages that produce the same hash value.
3) *Hadoop Distributed File System:* The Hadoop Distributed File System (HDFS) is designed for storing very large data sets reliably. In a big cluster, thousands of servers are able for both hosting directly attached storage and executing user applications. Hadoop cluster scales computation capacity, storage capacity, and I/O bandwidth up by simply adding commodity servers and distributing computation and storage among many servers.

VI. Privacy Analysis

A. Against Server's Attacks

To counter the attack from the semi-trusted servers, instead of using substitution cipher or k-anonymity approaches, we used MD5 algorithm. It is a cryptographically more secure and one-way function that produces random 128 bit digest values. Each item in the transaction database is converted into their corresponding digest values and no information regarding raw data is disclosed to the server. Only data owners keep the details of actual items but at the server side no information can be generated matching a specific digest value.

B. Against Data Owner's Attacks

Although data owners collaborate to mine their datasets but there are scenarios, where data owners are honest but curious to learn information of other data owners. To counter this attack in our solution, every data owner encrypts the customer IDs with RC4 encryption method using his private key. So no data owner is able to decrypt the customer details of other data owners. After getting the mining results from the server each data owner will decrypt that and get the original results.

VII. Experimental Setup And Results

The experiments are performed to compare the number of rules generated from original dataset by applying Apriori algorithm and that of from outsourced dataset through collaborative rule mining. The implementation has been done on Ubuntu 14.04. We have performed testing on market dataset of Yahoo. We have performed multiple runs by applying horizontal sampling on the dataset and taking different values of min sup and min conf. Final results are presented in the table III and IV.

Table III: Comparison of no of Rules Generated From Apriori Algorithm And Collaborative Rule Mining With Min_sup (2%) And Min_conf (50%)

Apriori Rule Mining				Collaborative Rule Mining		
Dataset	Rules	Dataset	Rules	Dataset1	Dataset2	Rules
Sample1	19	Sample2	16	Sample1	Sample2	65
Sample1	19	Sample3	58	Sample1	Sample3	84
Sample2	16	Sample3	58	Sample2	Sample3	104
Sample2	16	Sample4	31	Sample2	Sample4	83
Sample3	58	Sample4	31	Sample3	Sample4	77
Sample3	58	Sample5	49	Sample3	Sample5	122
Sample4	31	Sample5	49	Sample4	Sample5	118
Sample4	31	Sample6	17	Sample4	Sample6	86
Sample5	49	Sample6	17	Sample5	Sample6	101
Sample5	49	sample7	28	Sample5	Sample7	83
Sample6	17	Sample7	28	Sample6	Sample7	81
Sample7	28	Sample8	29	Sample7	Sample8	66
Sample8	29	Sample9	29	Sample8	Sample9	72
Sample9	29	Sample10	55	Sample9	Sample10	94

Table IV: Comparison of no of Rules Generated From Apriori Algorithm And Collaborative Rule Mining With Min_sup (2%) And Min_conf (60%)

Apriori Rule Mining				Collaborative Rule Mining		
Dataset	Rules	Dataset	Rules	Dataset1	Dataset2	Rules
Sample1	472	Sample2	469	Sample1	Sample2	911
Sample1	472	Sample3	471	Sample1	Sample3	1264
Sample2	469	Sample3	471	Sample2	Sample3	1045
Sample2	469	Sample4	641	Sample2	Sample4	1363
Sample3	471	Sample4	641	Sample3	Sample4	1113
Sample3	471	Sample5	446	Sample3	Sample5	1016
Sample4	641	Sample5	446	Sample4	Sample5	1315
Sample4	641	Sample6	462	Sample4	Sample6	1188
Sample5	446	Sample6	461	Sample5	Sample6	1215
Sample5	446	sample7	309	Sample5	Sample7	1140
Sample6	462	Sample7	309	Sample6	Sample7	1011
Sample7	309	Sample8	463	Sample7	Sample8	979
Sample8	463	Sample9	446	Sample8	Sample9	1204
Sample9	446	Sample10	537	Sample9	Sample10	1288

VIII. Conclusion

In This Paper, We Have Proposed A Privacy-Preserving Solution For Collaborative Rule Mining From Outsourced Data. This Allows Multiple Data Owners To Outsource Their Data Mining Jobs In A Privacy-Preserving Manner. Our Solution Protects Raw Data Of Participating Data Owners From The Third Party As Well As The Other data owners. Along with privacy-preservation, performance is also improved by using MapReduce framework. Our solution is suitable for the multi party environment, where high-level of privacy without compromising performance is required. Moreover, we also plan to explore some other important data mining techniques under the outsourced model and to integrate them with the proposed approach.

IX. Acknowledgment

The authors would like to thank Computer Engineering Department of Shri Govindram Seksaria Institute of Technology and Science, Indore for their motivation and support and for allocating the resources for the realization of the project.

References

[1] J. Vaidya, and C. Clifton, "Privacy-preserving data mining: Why, how, and when," *IEEE Security and Privacy,* vol. 2, no. 6, 1927, 2004.

[2] C. H. Tai, P. S. Yu, and D. N. Yang et al., "Privacy-preserving social network publication against friendship attacks," *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11),* pp. 1262-1270, 2011.

[3] E. Bertino, B. C. Ooi, Y. Yang, and R. H. Deng, "Privacy and ownership preserving of outsourced medical data," *Proc. of the 21st International Conference on Data Engineering (ICDE '05),* pp. 521-532, 2005.

[4] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Privacy in dynamic social networks," *Proc. of the 19th International Conference on World Wide Web (WWW 10),* pp. 1059-1060, 2010.

[5] K. Chard, K. Bubendorfer, S. Caton, and O. F. Rana, "Social cloud computing: A vision for socially motivated resource sharing," *IEEE Transactions on Services Computing,* vol. 5, no. 4, pp. 551-563, 2012.

[6] S. Pallickara, J. Ekanayake, and G. Fox, "A scalable approach for the secure and authorized tracking of the availability of entities in distributed systems," *Parallel and Distributed Processing Symposium (IPDPS '07),* 1-10, 2007.

[7] R. Aggarwal, A. Evmievski, and R. Srikant, Information sharing across private databases. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Database,* pp. 8697, San Diego, California, 2003.

[8] Agrawal, S. and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," In *Proceedings of the 21th IEEE International Conference on Data Engineering (ICDE 2005),* pp. 193-204, Tokyo, Japan, 2005.

[9] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "Randomdata perturbation techniques and privacy-preserving data mining," *Knowledge and Information Systems: An International Journal,* vol. 7, no. 4, pp. 387-414, 2005.

[10] L. Qiu, Y. Li, and X. Wu, "An approach to outsourcing data mining tasks while protecting business intelligence and customer privacy," In *Work-shops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006),* pp. 551-558, Hong Kong, China, December 1822, 2006.

[11] L. Qiu, Y. Li, and X. Wu, "Protecting business intelligence and customer privacy while outsouring data mining tasks," *Knowledge and Information Systems: An International Journal,* Nov. 16, 2007.

[12] B. Rozenberg, and E. Gudes, "Association rules mining in vertically partitioned databases," *Data Knowl. Eng.,* vol. 59, no. 2, pp. 378-396, 2006.

[13] J. Zhan, S. Matwin, and L. Chang, "Privacy-preserving collaborative association rule mining," In *Proc. DBSEC,* pp. 153-165, 2005.

[14] S. Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets, Inf. Sci., vol. 177, no. 2, pp. 490503, 2007.

[15] J. Vaidya, and C. Clifton, "Secure set intersection cardinality with application to association rule mining," *J. Comput. Secur.,* vol. 13, no. 4, pp. 593-622, 2005.

[16] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," In *Proc. SEDM,* pp. 345350, Jun. 2010.

[17] R. Kharat, M. Kumbhar, and P. Bhamre, "Efficient privacy preserving distributed association rule mining protocol based on random number," In *Intelligent Computing, Networking, and Informatics,* pp. 827836, Raipur, Chhattisgarh, India: Springer, 2014.

[18] K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining," In *Proc. VLDB,* 111-122, 2007.

[19] I. Molloy, N. Li, and T. Li, "On the (in)security and (im) practicality of outsourcing precise association rule mining," In *Proc. ICDM,* Dec. 2009, pp. 872-877

[20] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and W. Wang, "Privacy-preserving data mining from outsourced databases," in *Proc. CPDP,* 2011, pp. 411426.

[21] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving mining of association rules from outsourced transaction databases," *IEEE Syst. J.,* vol. 7, no. 3, pp. 385395, Sep. 2013.

[22] B. Zhou, J. Pei, "Preserving privacy in social networks against neighborhood attacks," *Proc. 24th International Conference on Data Engineering (ICDE' 08),* pp. 506-515, 2008.

[23] G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," *Knowledge discovery in databases*, pp. 229-238, 1991.

[24] J. Pei, "Preserving privacy in social networks against neighborhood attacks," *Proc. 24th International Conference on Data Engineering (ICDE '08),* pp. 506-515, 2008.