PRIVACY PRESERVATION IN DATA MINING USING ELLIPTICAL CURVE CRYPTOGRAPHY

Krunal K Patel[1], Asst Prof. Risha Tiwari[2]
[1]P.G Student, [2]Assistant Professor, Department of Computer Engineering
Hasmukh Goswami College of Engineering, Ahmedabad, India

Abstract: There are many distributed and centralized data mining techniques often used for various applications. Privacy and security issues of these techniques are recently investigated with a conclusion that they reveal information or data to each other clients involved to find global valid results. But because of privacy issues, involving clients do not want to share such type of data. Recently many cryptography algorithms have been found to address privacy problems in distributed and centralized data mining. In this thesis, we propose an elliptic curve cryptography based algorithm to mine privacy-preserving association rules on horizontally partitioned data. Elliptic Curve Integration Encryption Scheme is used for security of data and Elliptic Curve Digital Signature Algorithm for authentication. Moreover, we have also considered unsecured communication channels in distributed environment. Proposed algorithm provides privacy and security against involving clients and other clients (adversaries) who can reveal information by reading unsecured channel between involving clients. Finally, we analyse the privacy and security provided by proposed algorithm.

Keywords: Privacy preservation, frequent item set, rule mining, elliptic curve cryptography, ECIES, ECDSA

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.[12] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [12]. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness [1]. Association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onions, potatoes} => {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.

II. RELATED WORK

A. Privacy Preserving Mining of Association Rules on Horizontally and Vertically Partitioned Data

The methods proposed in three papers based on Privacy-Preserving Data Mining PPARM in horizontal partitioning of databases. The security of the scheme heavily draws on the computational difficulty of discrete logarithm problem (DLP). Enhanced M.Hussein et al.'s Scheme EMHS does not modify the original data in both two phases. Thus, Initiator will find global frequent item sets accurately means the final results are accurate [1]. Commutative encryption is used for algorithm proposed in which didn’t violate privacy constraints. EMHS scheme satisfies semi-honest model. EMHS uses Paillier cryptosystem in the second phase and MHS uses RSA cryptosystem. For this reason, Combiner is much more difficult to attack in EMHS. Means EMHS has higher privacy than MHS. Both EMHS and MHS are two phase schemes, the communication cost (or cost) of each scheme is the sum of the one in each phase. EMHS has better performance than MHS in sparse datasets when increasing the number of sites.

B. Elliptic Curve Cryptography Based Mining of Privacy Preserving Association Rules in Unsecured Distributed Environment

Elliptic curve provides public cryptosystem based on discrete logarithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide security level same as RSA with larger key length. A detailed overview of elliptic curve and elliptic curve cryptosystem is given in. Chirag Modi, Udai Pratap Rao and

Dhiren R. Patel used elliptic curve based Diffie-Hellman (ECDH) protocol and elliptic curve based Digital Signature algorithm (ECDSA) for key exchange and authentication-verification respectively in our proposed algorithm. In following we give an overview of ECDH protocol and ECDSA.

Privacy against Involving Sites and TP: In proposed communication protocol each site sends only local support count of items to next site instead of original transaction data. So, next site or any other site cannot be able to know the actual contents of transactions. Moreover each site sends information to next site after encrypting it by using the shared key with TP. So any other site cannot predict the original value. TP also cannot predict local support of an item set contained at any site because it gets total encrypted support count of item set from the last site[3].

C. Fast Private Association Rule Mining by A Protocol for Securely Sharing Distributed Data

In this research paper proposed a flexible and easy-to-implement protocol for privacy preserving data sharing based on a public-key crypto-system. The protocol is efficient in practical settings and it requires less machinery than previous approaches (where commutative encryption was required). The sharing process was reduced from 3p steps to 2p steps. Commutative encryption can be implemented with RSA; however, all schemes are not only more expensive that proposed here, but commutative encryption has a probability/certainty factor that requires delicate management [4]. These methods are safe. Thus, methods ensure that no data (transactions) can be linked to a specific user. The protocol allows users to conduct private mining analyses without loss of accuracy (as opposed to data sanitization). Protocol works under the common and realistic assumption that parties are semi-honest, or honest but curious, meaning they execute the protocol exactly as specified, but they may attempt to infer hidden links and useful information about other parties. The requirement for non-collusion is higher than previous methods, since p - 1parties must collude to compromise the protocol. Vladimir Estivill-Castro and Ahmed Haj Yasien leak some information, but this is less than what previous methods have regarded as innocuous and previous research has explored whether parties are willing to trade off the benefits and costs of sharing sensitive data. The results of this research showed that parties are willing to trade-off privacy concerns for economic benefits. There are few issues that may influence practical usage of the presented protocol. The issue of having a party that can be trusted with shuffling the records and publishing the database to all other parties can be solved with slightly more sophistication, but essentially no overhead; if there are p parties, each party plays the data distributor withl/p share of the data, and we conduct p parallel rounds.

D. Prospective Utilization of Elliptic Curve Cryptography for Security Enhancement

Although RSA, El-GAMAL and Diffie –Hellman are secure asymmetric key cryptosystem, their security comes with a price ,their large keys. So researchers have looked for providing substitute that provides the same level of security with smaller keys. Efficiency of ECC is depends upon factors such as computational overheads ,key size, bandwidth ,ECC provides higher-strength per- bit which include higher speeds, lower power consumption, bandwidth savings, storage efficiencies, and smaller certificates [5].

E. Securing User’s Data in Hadoop Distributed File Sytem (HDFS)

Scenarios for Experiments Basically Hetalben Gajjar created a cluster of three nodes with hardware and software configurations mentioned in paper. Out of three nodes one act as Name Node, Secondary Name Node and Data Node itself and the other two of the three are given roles of Data Nodes only. So all together we have three Data Nodes. In Scenario 1 the configuration of replication factor is 2 that is two replicas are stored on Data Nodes in HDFS for each file. While in scenario 2 the set up is same but the replication factor is set to 3 so that three instances of each file are stored on the Data Nodes which makes it more robust. We have not considered the replication factor 1 as only single instance of each file is stored so the system will not be considered reliable. The time taken to write a file to HDFS in all the four cases is measured by copying file from local system that is running as a Virtual Machine on same host to HDFS. Similarly the read time is measured in all four cases by copying the files stored on HDFS to local system [6]. Approach is based on Elliptic Curve Integrated Encryption System to harness the Hadoop Distributed File System with security. In addition to provisioning for data confidentiality this implementation also provides integrity of user’s data. Also a new random secret key is generated for each file that is stored on HDFS. However the user is freed from the overhead of secret key management as it is transparent to user.

F. Achieving Authentication and Integrity using Elliptic Curve Cryptography Architecture

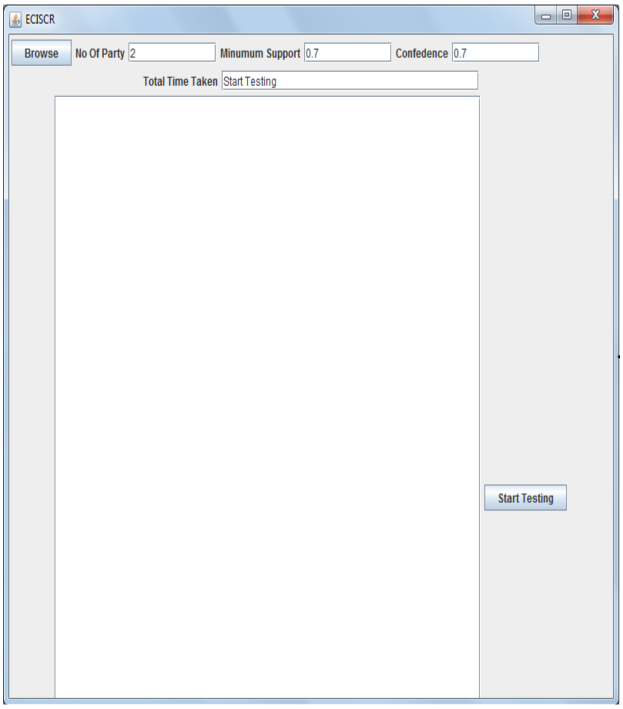
Cost for creating session keys the sender generates two elliptic curve points: 1). which derives the symmetric encryption and the MAC key MACk1. 2). which is used by the receiver to derive the point P. Generation of these two points require two elliptic curve scalar multiplication procedure, which takes 2log(p). The receiver calculates the point P from the request message and the cost is incurred is log(p) [8]. Cost for encryption messages :The cost of encrypting messages using ECIES is the cost incurred in encrypting AES in CBC mode, which is certainly less than the cost of calculating keys for RSA.

III. IMPLEMENTATION OF ECIES AND ECDSA ALGORITHMS

We have implemented ECIES and ECDSA Algorithms using Java. Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java, Eclipse can be used to develop applications. It can also be used to develop packages for the software Mathematical. Development environments include the Eclipse Java development tools (JDT) for Java and Scala, Eclipse CDT for C/C++ and Eclipse PDT for PHP, among others. The file that used java class file. Java class file is a file (with the .class filename extension) containing a Java byte code which can be executed on the Java Virtual Machine (JVM). Java class file is produced by Java compiler from Java programming language source files (.java files) containing Java classes. If a source file has more than one class, each class is compiled into a separate class file.

IV. RESULT OF IMPLEMENTATION WORK AND DISCUSSION

Here As given in the screen we have to give input values as shown in figure 1.We have to give one Data file, Number of Parties, minimum support value and minimum confidence value as input to system. Here input file is DAT file and all the values are separated by space it contains row wise data. As we give input it will divide the data base horizontally with number of parties and assign that many transactions to each party. Now start with any party as initiator and count local frequency set of that party using Hash based Apriori Algorithm and this count is added to any random number that will be encrypted by ECIES algorithm and signed by ECDSA algorithm now that will go to another party will be decrypted by that party and find his local frequency set using Hash based Apriori and thus encrypt data as explained above and also decryption of another user so at last in encrypted data will be arrived at initiator and initiator will be decrypt data and remove the random number that was added at initialization time.

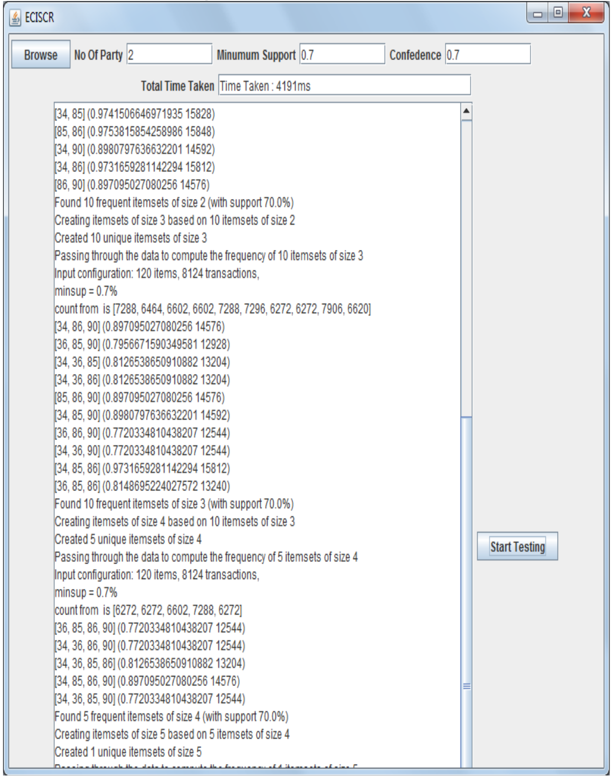


No Of Party 2
Minumum Support 0.7
Confedence 0.7
Total Time Taken Start Testing

Figure 1 input Front end of system

As this will calculate values for size of one then combining values in two, three and so on up to when there is at least one transaction found in scanning as output.

Now when output will be generated it will be display in screen as shown in figure 2.



No Of Party 2
Minumum Support 0.7
Confedence 0.7
Total Time Taken Time Taken : 4191ms

```
[34, 85] (0.9741506646971935 15828)
[85, 86] (0.9753815854258986 15848)
[34, 90] (0.8980797636632201 14592)
[34, 86] (0.9731659281142294 15812)
[86, 90] (0.897095027080256 14576)
Found 10 frequent itemsets of size 2 (with support 70.0%)
Creating itemsets of size 3 based on 10 itemsets of size 2
Created 10 unique itemsets of size 3
Passing through the data to compute the frequency of 10 itemsets of size 3
Input configuration: 120 items, 8124 transactions,
minsup = 0.7%
count from  is [7288, 6464, 6602, 6602, 7288, 7296, 6272, 6272, 7906, 6620]
[34, 86, 90] (0.897095027080256 14576)
[36, 85, 90] (0.7956671590349581 12928)
[34, 36, 85] (0.8126538650910882 13204)
[34, 36, 86] (0.8126538650910882 13204)
[85, 86, 90] (0.897095027080256 14576)
[34, 85, 90] (0.8980797636632201 14592)
[36, 86, 90] (0.7720334810438207 12544)
[34, 36, 90] (0.7720334810438207 12544)
[34, 85, 86] (0.9731659281142294 15812)
[36, 85, 86] (0.8148695224027572 13240)
Found 10 frequent itemsets of size 3 (with support 70.0%)
Creating itemsets of size 4 based on 10 itemsets of size 3
Created 5 unique itemsets of size 4
Passing through the data to compute the frequency of 5 itemsets of size 4
Input configuration: 120 items, 8124 transactions,
minsup = 0.7%
count from  is [6272, 6272, 6602, 7288, 6272]
[36, 85, 86, 90] (0.7720334810438207 12544)
[34, 36, 86, 90] (0.7720334810438207 12544)
[34, 36, 85, 86] (0.8126538650910882 13204)
[34, 85, 86, 90] (0.897095027080256 14576)
[34, 36, 85, 90] (0.7720334810438207 12544)
Found 5 frequent itemsets of size 4 (with support 70.0%)
Creating itemsets of size 5 based on 5 itemsets of size 4
Created 1 unique itemsets of size 5
```

Figure 2 output screen

V. CONCLUSION

For Data mining there are lots of algorithms and also to provide security and privacy of data. Weanalysed security and privacy of proposed algorithm against involving sites or adversary. Here elliptic curve integrated encryption scheme is used for encryption and decryption and elliptic curve digital signature algorithm is used to find where involving user is authentic or not. Elliptic curve integrated encryption scheme provides better security than RSA algorithm with less key length. So it will improve speed of encryption-decryption and also use less space. In this scheme any database can act as initiator and count global frequency set and generates global association rules.

VI. ACKNOWLEDGMENT

I wish to express my deep sense of gratitude to my guide, Prof. Risha Tiwari, Assistant Professor, computer engineering and Prof. Indrjeet Rajput, Assistant Professor, computer engineering for their guidance and useful suggestions, which helped me to complete my research paper. Finally yet importantly, I would like to thank my beloved parents for their blessing. My friends and classmate for their help and wishes for successful completion of this research paper.

REFERENCES

[1] MadhuriN.Kumbhar and Ms.ReenaKharat, “Privacy Preserving Mining of Association Rules on Horizontally and Vertically Partitioned Data: A Review Paper”, International Conference onHybrid Intelligent Systems, 2012 IEEE.

[2] Ashraf B. El-Sisi and Hamdy M. Mousa, “Evaluation of Encryption Algorithms for Privacy Preserving Association Rules Mining”,International Journal of Network Security, Vol.14, No.5, PP.284-291, Sept. 2012

[3] Chirag N. Modi, UdaiPratapRao and Dhiren R. Patel, “Elliptic Curve Cryptography Based Mining of Privacy Preserving Association Rulesin Unsecured Distributed Environment”, International Conference on Advances in Communication, Network, and Computing, 2010 IEEE

[4] Vladimir Estivill-Castro and Ahmed HajYasien, “Fast Private Association Rule Mining by A Protocolfor Securely Sharing Distributed Data”, 2007 IEEE.

[5] SonaliNimbhorkar and Dr.L.G.Malik, “Prospective Utilization of Elliptic Curve Cryptography for Security Enhancement”, International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 1, January 2013

[6] HetalbenGajjar, “Securing User’s Data in HDFS”, International Journal of Computer Trends and Technology (IJCTT) - volume4 Issue 5–May 2013

[7] RuchikaMarkan and GurvinderKaur, “Literature Survey on Elliptic Curve Encryption Techniques”, International Journal of Advanced Research inComputer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 9, September 2013

[8] Ms.ManaliDubal and Ms.AaradhanaDeshmukh, “Achieving Authentication and Integrity using Elliptic Curve Cryptography Architecture”, International Journal of Computer Applications,Volume 69– No.24, May 2013.

[9] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail, “Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous DataBase”, Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607 -- 616 (2008).

[10] MoezWaddey , Pascal Poncelet, Sadok Ben Yahia, “Novel Approach For Privacy Mining Of Generic Basic Association Rules,” In PAVLAD’09, November 6, 2009, Hong Kong, China, 2009 ACM.

[11] XuanCanh Nguyen, HoaiBac Le, Tung Anh Cao, “An Enhanced Scheme For Privacy-Preserving Association Rules Mining On Horizontally Distributed Databases,” In 2012 IEEE.

[12] http://en.wikipedia.org/wiki/Data_mining