Scientific Journal of Impact Factor (SJIF): 3.134

ISSN (Online): 2348-4470
ISSN (Print) : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 4, April -2015

SECURED MULTIPARTY DATA FUSION AND EXTRACTION OF PRIVATE DATA

Roja.M[1], Mr.S.P.Rajagopalan[2]

[1] *PG Scholar, M.E Software Engineering, GKM college of Engineering and Technology, Chennai,.*

[2] *Professor, Department of Computer Science and Engineering, GKM college of Engineering and Technology, Chennai.*

Abstract- *Privacy-preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information. The problem of private data publishing, where different attributes for the same set of individuals are held by two parties is addressed by a two party authentication algorithm. In particular, an algorithm for differentially private data release for vertically partitioned data between two parties in their semi-honest adversary model is proposed. The proposed algorithm is applied to the banking system where private data of the customers are required for banking and for the data maintenance. In the banking system there is no differentiability in the private data each data will be processed and retrieved from the distinct data base only. The data will be processed from two distinct and different databases. Each data has unique link between them using that link the data will be retrieved. Exraction of New Data from the Merged Data is performed. The implementation is all about Company Employee who has got Loan. Employee ID plays as Primary Key and the List of Loan obtainers can be identified. Data is analyzed only by the Authorized Persons. The security of the private data of the employee transferred between the company and the bank is enhanced and the results are stored in the data base.*

Keywords*-Semi-honest adversary, vertically partitioned, Quasi-identifier (QID), privacy preserving data mining, multi party computation.*

I. INTRODUCTION

Huge databases exist today due to the rapid advances in communication and storing systems. Each database is owned by a particular autonomous entity, for example, medical data by hospitals, income data by tax agencies, financial data by banks, and census data by statistical agencies. The data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time, new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration. An algorithm to securely integrate person-specific sensitive data from two data providers is proposed, whereby the integrated data still retain the essential information for supporting data mining tasks. This research problem was discovered in a collaborative project with the financial industry. The single-party algorithm for differential privacy that has been recently proposed is taken into account. So at any time during the execution of the algorithm, no party should learn more information about the other party's data than what is found in the final integrated table, which is differentially private. We present a two-party protocol for the exponential mechanism. We use this protocol as a sub protocol of our main algorithm, and it can also be used by any other algorithm that uses the exponential mechanism in a distributed setting with use of big data base. In this paper an algorithm for differentially private data release for vertically partitioned data between two parties in the semi honest adversary model is used. Also it uses various protocols to retrieve the data. Extraction of new data from the merged data is performed. The Implementation is all about Company Employee who has got Loan. Employee ID plays as Primary Key and we can identify the List of Loan obtainers. Data is analyzed only by the Authorized Persons.

1.1. Problem statement

The problem of private data publishing, where different attributes for the same set of individuals are held by two parties is addressed. In particular, an algorithm for differentially private data release for vertically partitioned data between two parties in the semi honest adversary model is presented an implemented in a banking system. Avoids forgery of loan originators and Privacy-preserving data publishing actions.

1.2. Objectives

A heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data (HACE) theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective is used. This data-driven model involves demand driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data driven model and also in the Big Data revolution. We propose a two-party algorithm that releases differentially private data in a secure way according to the definition of

secure multiparty computation. Experimental results on real-life data suggest that the proposed algorithm can effectively preserve information for a Data task.

1.3. Scope

The proposed paper improves the privacy-preservation using a two party authentication algorithm. The scope is to provide loan to qualified employees by the corresponding bank. Data is analyzed only by the Authorized Persons that is company and bank administration. It achieves Fast retrieval of data from big data base.

1.4. Issues

There is no differentiability in the private data each data will be processed and retrieved from the distinct data base only. It provides similar data utility compared to the recently proposed single-party algorithm and better data utility than the distributed k-anonymity algorithm for classification analysis.

II. EXISTING PRIVACY MODEL

2.1. Randomization

Randomization technique is an inexpensive and efficient approach for privacy preserving data mining (PPDM). In order to assure the performance of data mining and to preserve individual privacy, this randomization schemes need to be implemented. The randomization approach protects the customers' data by letting them arbitrarily alter their records before sharing them, taking away some true information and introducing some noise. Some methods in randomization are numerical randomization and item set randomization Noise can be introduced either by adding or multiplying random values to numerical records or by deleting real items and adding fake values to the set of attributes.

2.2. Anonymization

To protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity. In order to preserve privacy, k-anonymity model has been proposed by Sweeney which achieves k-anonymity using generalization and suppression, In K-anonymity, it is difficult for an imposter to determine the identity of the individuals in collection of dataset containing personal information. Each release of data contains every combination of values of quasi-identifiers and that is indistinctly matched to at least k-1 respondents.

2.2.1. Generalization

Generalization involves replacing a value with a less specific generalized but semantically reliable value. For example, the age of the person could be generalized to a range such as youth, middle age and adult without specifying appropriately, so as to reduce the risk of identification.

2.2.2. Suppression

Suppression involves reduce the exactness of applications and it does not liberate any information by using this method it reduces the risk of detecting exact information. It protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity.

2.3. K-Anonymity

It presents a two-party framework along with an application that generates k-anonymous data from two vertically partitioned sources without disclosing data from one site to the other. A new multidimensional model, which provides an additional degree of flexibility, is proposed. Here it introduces a simple greedy approximation algorithm, and experimental results show that this greedy algorithm frequently leads to more desirable anonymization than exhaustive optimal algorithms for two single-dimensional models. It proposes a K-anonymization solution for classification. A useful approach to combat linking attacks called K-anonymization.it is the process of anonymzing the linking attributes so that at least k released records match each value combination of the linking attributes. The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. A dataset complies with k-anonymity protection if each individual's record stored in the released dataset cannot be distinguished from at least k-1 individuals whose data also appears in the dataset.

2.3.1. Quasi-identifier

Quasi-identifier (QID) is a set of features whose associated values may be useful for linking with another data set to re-identify the entity that is the subject of the data. While releasing private tables for research purpose identifiers are removed from the table to de-identify the person but still by matching quasi-identifiers from private table with public table one can easily identify the person. Therefore k-Anonymization is used to make at least k tuples similar by using

generalization or suppression. A set of decentralized protocols that enable data sharing for horizontally partitioned databases. Here it includes a new notion, l-site-diversity, for data anonymization to ensure anonymity of data providers. A new privacy model called LKC-privacy to overcome the challenges and present two anonymization algorithms to achieve LKC-privacy in both the centralized and the distributed scenarios. It develops a data publishing technique that ensures ϵ -differential privacy while providing accurate answers for count queries where the predicate on each attribute is a range. It proposes the first anonymization algorithm for the non-interactive setting based on the generalization technique. It probabilistically generalizes the raw data and then adds noise to guarantee $\in$-differential privacy. It proposes two algorithms to securely integrate private data from multiple parties (data providers).The first algorithm achieves the k-anonymity privacy model in a semi-honest adversary model. The second algorithm employs a game-theoretic approach to thwart malicious participants and to ensure fair and honest participation of multiple data providers in the data integration process.

2.4. Privacy Preservation

Information system must persuade one of the most important properties as Privacy. For this basis, several efforts have been dedicated to incorporating privacy preserving techniques with data mining algorithms in order to prevent the revelation of sensitive information during the knowledge finding. Existing privacy preserving data mining techniques can be classified according to the following five different Dimensions the modification applied to the data (perturbation, substitution, generalization, encryption and so on) in order to sanitize data distribution (centralized or distributed)
the data type (single data items or complex data correlations) that needs to be protected from disclosure the data mining algorithm which the privacy preservation technique is designed for heuristic or cryptography-based approaches Crypto graphy-based algorithms. Cryptography-based algorithms are designed for protecting privacy in a distributed scenario by using encryption techniques while heuristic based techniques are mainly conceived for centralized datasets,. Heuristic-based algorithms just projected aim at defeat sensitive raw data by applying perturbation techniques based on probability distributions. Furthermore, several heuristic-based approaches for hiding both raw and aggregated data through a hiding techniques (k-anonymization, adding noises, data swapping, generalization and sampling) have been developed, first, in the context of association rule mining and classification and, more recently, for clustering techniques.

2.5. Two Party Privacy Authentications

Although generic constructions for secure computation can, in principle, efficiently compute any polynomial function, the resulting overhead is often unacceptable. This might be due to the size of the circuit computing the function, or to the fact that each input value (or sometimes, as in the two-party case, each input bit) incurs expensive operations such as input sharing or computing an oblivious transfer. In general, when considering semi-honest adversaries and a reasonably sized circuit, the protocols are reasonably efficient. However, when considering malicious adversaries these protocols are typically not practical even for small circuits. Three specialized constructions which are considerably more efficient than applying generic constructions to the same functions. The constructions are secure against semi-honest adversaries, although for some of them there exist variants which are secure against malicious adversaries. The constructions are based on the use of homomorphism encryption, oblivious polynomial evaluation, and the reduction of the computed function to simpler functionalities, with the analysis of the resulting protocol in the hybrid model .For each function, we describe the overhead of applying a generic construction, then describe the basic details of the specialized construction and its overhead.

III. SECURE MULTI PARTY COMPUTATION

The increasing use of data mining tools in both the public and private sectors raises concerns regarding the potentially sensitive nature of much of the data being mined. The utility to be gained from widespread data mining seems to come into direct conflict with an individual's need and right to privacy. Privacy preserving data mining solutions aims at achieving the somewhat paradoxical property of enabling a data mining algorithm and to use data without ever actually seeing it. Thus, the benefits of data mining can be enjoyed, without compromising the privacy of concerned individuals. The aim of a secure multiparty computation task is for the participating parties to securely compute some function of their distributed and private inputs. A key question that arises here is what it means for a computation to be secure. One way of approaching this question is to provide a list of security properties that should be preserved. The first such property that often comes to mind is that of privacy or confidentiality. A naïve attempt at formalizing privacy would be to require that each party learns nothing about the other parties' inputs, even if it behaves maliciously. However, such a definition is usually unattainable because the defined output of the computation typically reveals some information on other parties' inputs. For example, a decision tree computed on two distributed databases reveals some information about both databases. Therefore, the privacy requirement is usually formalized by saying that the only information learned by the parties in the computation again, even by those who behave maliciously is that specified by the function output. Although privacy is a primary security property, it rarely suffices. Another important property is that of correctness; this states that the parties' output is really that defined by the function if correctness is not guaranteed, then a malicious party may be able to receive the specified decision tree while the honest party receives a tree that is modified to provide

misleading information. A central question that arises in this process of defining security properties is: when is our list of properties complete. The question is, of course, application-dependent and this essentially means that for every new problem, the process of deciding which security properties are required must be re-evaluated.

A. Architecture

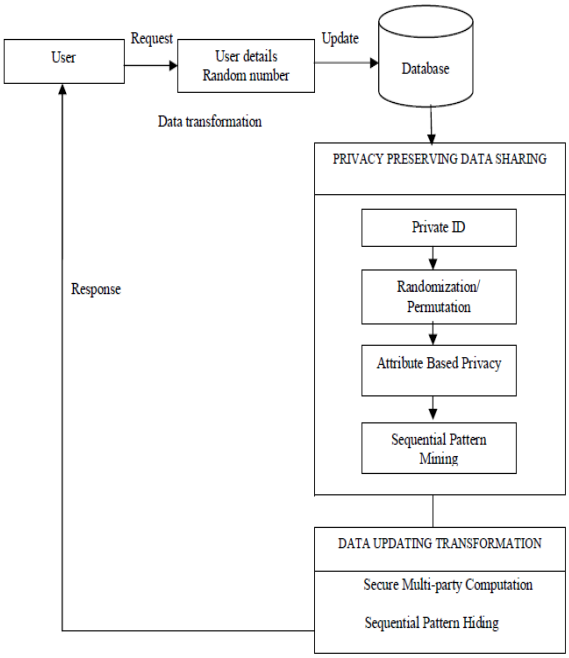

Fig. 1.1 Secure multi party computation

B. Multi Party Authentication Algorithm

Steps

Let P1 and P2 be the parties and let I denote the indices of the corrupted parties controlled by an adversary A. In principle, it is possible for zero, one, or both parties to be corrupted. However, for the sake of simplicity, we will consider the most important case, that either I = {1} or I = {2} (i.e., exactly one of the two parties is corrupted). An ideal execution proceeds as follows:

Input

Each party obtains an input; the ith party's input is denoted xi. The adversary A receives an auxiliary input denoted z.

Step 1: Send inputs to trusted party

The honest party Pj for j /∈ I send its input xj to the trusted party. The corrupted party Pi for i ∈ I (who is controlled by A) may abort by replacing the input xi with a special abort message, send its input x_i, or send some other input of the same length to the trusted party. This decision is made by A and may depend on the value xi for i ∈ I and its auxiliary input z. Denote the inputs sent to the trusted party by (w1, w2) (note that w_i does not necessarily equal xi).

If the trusted party receives an input of the form abort from Pi, it sends abort to both parties and the ideal execution terminates. Otherwise, the execution proceeds to the Next step.

Step 2: Trusted party sends outputs to adversary

The trusted party computes the pair of outputs (f1 (w1, w2), f2 (w1, w2)) and sends fi (w1, w2) to the corrupted party Pi.

Step 3: Adversary instructs trusted party to continue or halt

A sends either continue or abort to the trusted party. If it sends continue, the trusted party sends fj (w1, w2) to the honest party Pj. Otherwise, if A sends abort, the trusted party sends abort to party Pj.

Output

The honest party always outputs the message it obtained from the trusted party. The corrupted party outputs nothing. The adversary outputs any arbitrary (probabilistic polynomial-time computable) function of the initial input xi, the auxiliary input z, and the output abort or fi (w1, w2) obtained from the trusted party.

IV. RELATED WORK

All the User details will be stored in the Database of the Service Provider. Once the User creates an account, they are to login into their account and request the Job from the Service Provider. Based on the User's request, the Service Provider will process the User requested Job and respond to them. The User information will be stored in the Database of the Company Service Provider. Company server will contain the large amount of data in their Data Storage. The Company Service provider will maintain the all the User information to authenticate when they want to login into their account. Also the Company Server will redirect the User requested job to the any of the Queue to process the User requested Job. The Request of all the Users will process by Company Server will establish connection between them. For this Purpose we are going to create an User Interface Frame. Also the Company Service Provider will send the User Job request to the Queues in Fist in First out manner. The Bank Service provider will maintain the all the User information to authenticate when they want to login into their account. The User information will be stored in the Database of the Bank Service Provider. Bank Service Provider will contain information about the user in their Data Storage. To communicate with the Client and with the other modules of the Company server, the Bank Server will establish connection between them. A concept of merged data that is in a company or organization will maintain the employee information both private and public data is implemented. The employee may contain private data like employee id, employee name , salary and the loan applied and loan go and public data like email id and phone number. But more private information like bank account number and pin number are not reveled form the company. So we merge the both type information into one new table. The two party authentications are done by both bank and company to list the log of the employee whether he is eligible to take up loan. so the authenticated by bank through the company and company will provide a set of information it will be validated by the both bank and company by using scheme of two party authentication. If user fails in this scheme he cannot take loan any service provided by the bank.

V. CONCLUSION

The proposed algorithm is differentially private and secure under the security definition of the semi honest adversary model. It can effectively retain essential information for classification analysis. It provides similar data utility compared to the recently proposed single-party algorithm and better data utility than a distributed k-anonymity algorithm for classification analysis. In the next phase the retrieved data will be merged for purpose of providing information to the bankers and further the database will be analyzed by the big data analytics, finally the result will be extracted from data base

VI. FUTURE WORK

The future is a pioneer work of the proposed technique to Big Data. A suitable Big Data tool like Hadoop or HDFS is chosen and the data mining of the proposed multi-party authentication is applied. This application in Big Data is efficient at a larger pace since all the upcoming data mining and data authentication is based large data sets.

REFERENCES

[1] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "MondrianMultidimensional K-Anonymity," Proc. IEEE Int'l Conf. DataEng. (ICDE '06), 2006.
[2] W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," Very Large Data Bases J., vol. 15, no. 4, pp. 316-333, Nov. 2006.
[3] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and DataEng., vol. 19, no. 5, pp. 711-725, May 2007.
[4] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-AwareAnonymization Techniques for Large-Scale Data Sets," ACMTrans. Database Systems, vol. 33, article 17, 2008.
[5] P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers," Proc. Ann. IFIP WG 11.3 Working Conf. Data and Applications Security (DB Sec '09), 2009.
[6] N. Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu, "Differentially Private Data Release for Data Mining," Proc. ACM Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '11), 2011.

[7] N. Mohammed, B.C.M. Fung, P.C.K. Hung, and C. Lee, "Centralizedand Distributed Anonymization for High-DimensionalHealthcare Data," ACM Trans. Knowledge Discovery from Data vol. 4, no. 4, pp. 18:1-18:33, Oct. 2010.
[8] A.Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private Record Matching Using Differential Privacy," Proc. Int'l Conf.Extending Database Technology (EDBT 10), 2010.
[9] Friedman and A. Schuster, "Data Mining with Differential Privacy," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '10), 2010.
[10] A.Narayan and A. Haeberlen, "DJoin: Differentially Private JoinQueries over Distributed Databases," Proc. 10th USENIX Conf.Operating Systems Design and Implementation (OSDI '12), 2012.

