IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

Sharing Private Data for Building Data Analysis Model

P Sathya[1], E Shanmugapriya[2], R Ranjani[3]

[1, 2, 3]**Department of Information Technology, Anand Institute of Higher Technology, Anna University, Chennai, Tamil Nadu, India.**

Abstract

In data analysis model, the collaborating parties with their private data make use of privacy preserving data analysis (PPDA) method to learn the analysis results. Here, different parties have different incentives. In this project each party participates in a protocol to learn the output of some function *f* over the joint inputs of the parties. The parties which are connected to PPDA send their inputs to PPDA which is also called as single data analysis module. Certain PPDA technique reveals only the final result. It is impossible to verify whether the participating parties provide their true private inputs to PPDA. Therefore, proper incentives must be set by the participating parties. Here, we aim to create privacy preserving data analysis model that motivate the participating parties to provide true inputs. Users can determine the outcome of the data analysis by the parameters they chose, thus providing additional value to business strategies and initiatives. It is important to note that without these parameters, the data mining program will generate all permutations or combinations irrespective of their relevance.

Data mining programs lack the human intuition to recognize the difference between a relevant and an irrelevant data correlation, users need to review the results of mining exercises to ensure results provide needed information. Privacy and security, particularly maintaining the confidentiality of data have become a challenging issue with advances in information and communication technology. The ability to communicate and share the data has many benefits too. So, certain privacy preserving data analysis tasks are analysed in such a way that telling the truth is the best choice for any participating party.

Index Terms*-PPDA, data analysis, privacy, security.*

I. Introduction

Recent advances in the data mining field have lead to increased concerns about privacy**.** The topic of privacy has been traditionally studied in the context of cryptography and information-hiding. Recent emphasis on data mining has lead to renewed interest in the field. In data analysis model, when the participating parties send their inputs to PPDA, the data can be misused. To prevent the misuse of data, several laws were proposed for providing confidentiality to the data. But this protection comes with a real cost which includes both added security cost and penalties. We need is the ability to compute the desired “beneficial outcome” of data sharing. Secure Multiparty Computation (SMC) emerged as an answer to this problem. But the drawback in SMC is that nothing other than the final results are revealed and the final result is based on the party’s own incentives. Also, SMC protocol does not guarantee that all the participating parties provide their true inputs. But we generally assume that participating parties provide true inputs. It is justified by the fact that learning the correct data analysis results or models is in the best interest of all participating parties. SMC-based protocols involve expensive computations. Therefore, any party that does not want to learn the analysis results, the party should not participate in the protocol. Still we cannot guarantee the truthfulness of the private input data when the participating parties want to learn the final results exclusively. Unless proper incentives are set, current SMC techniques cannot prevent input modifications by participating parties. In this paper, we assume that the number of malicious or dishonest participating parties are at most *n-1,* where *n* is the number of parties. This assumption is very general since most existing works in the area of privacy-preserving data analysis assume either all participating parties are honest or semi-honest or the majority of the participating parties are honest. We explore which functionalities can be implemented in a way that participating parties have the incentive to provide their true private inputs upon engaging in the corresponding SMC protocols. Thus, in this paper we use another method called Non Co-operative Computation (NCC) technique. We extend the non cooperative computation definitions to incorporate cases where there are multiple dishonest parties. In this paper, NCC technique is used in the existing system where it is explained in section 2. In addition, we show that from incentive compatibility point of view, most data analysis tasks need to be analyzed only for two party cases. Therefore, to overcome the drawbacks by SMC-based protocol, we make use of other methods and cryptographic techniques which provides security and privacy, and also confidentiality of the data which has to be maintained.

The rest of this paper is introduced as follows: In section 2, presents the existing work of the system regarding the non cooperative computation. In section 3, explanation regarding the proposed system is given. In section 4, brief explanation about the system architecture is provided and section 5, concludes this system and represents future work of this paper.

ABBREVIATIONS	EXPANSION OF ABBREVIATIONS
NCC	Non-cooperative computation
DNCC	Deterministic NCC
PPDA	Privacy Preserving Data Analysis
SMC	Secure Multi-party Computation
TTP	Trusted Third Party

Table 1 List of Terms used in this system.

II. Related Work

Privacy Preserving Data Analysis techniques cannot prevent participating parties from modifying their private inputs. It is a very complex and time-consuming task that requires an extraordinary design and implementation efforts from controlling the participating parties to change their private incentives. In the existing system, we have prior knowledge about the schema, structure, vocabulary or any technical details of these sources. All these details are collected by the PPDA module. Many privacy-preserving data analysis techniques have been designed using cryptographic techniques. Data are generally assumed to be either vertically or horizontally partitioned. In PPDA we assume the parties are semi-honest in sharing the information. In the existing system, we make use of the NCC technique. NCC is used for parties that do not co-operate in computation. The NCC technique makes the following assumptions:

1. Correctness. The first priority for every participating party is to learn the correct result.

2. Exclusiveness. If possible, every participating party prefers to learn the correct result exclusively.

The concept of non-cooperative computation (NCC), which is the joint computation of a function by self-motivated agents, where each of the agents possesses one of the inputs to the function. In NCC the agents communicate their input (truthfully or not) to a trusted center which performs a commonly-known computation and distributes the results to the agents. NCC is a game theoretic concept.

NCC maintains two things. 1) Cryptography techniques 2) Trusted Third Party (TTP). The inputs from parties are sent to PPDA. Then, these incentives are sent from PPDA to TTP to check whether the inputs sent by the respective parties are true inputs. If the inputs are true, then the TTP sends positive acknowledgement to PPDA. Otherwise, negative acknowledgement is sent to PPDA by TTP. Then, Privacy preserving data analysis module responses to the requested parties with the resource. Trusted Third Party (TTP) resolves the conflicts between two parties. It also checks the inputs.

NCC is a very broad framework, and is specialized by imposing specie structure on the agents' utility functions. The technical results we present are specie to the setting in which each agent has a primary interest in computing the function, and a secondary interest in preventing the others from computing it (properties called correctness and exclusiveness).

Here, the PPDA is called the single data analysis module which stores the inputs sent by the participating parties. The input is sent to PPDA in the form of query.ie, the organization's names along with the inputs are sent as input to single data analysis module. By using the name of the party, TTP checks the respective inputs of the parties.

In cryptography, a trusted third party (TTP) is an entity which facilitates interactions between two parties who both trust the third party; The Third Party reviews all critical transaction communications between the parties, based on the ease of creating fraudulent digital content. In TTP models, the relying parties use this trust to secure their own interactions. TTPs are common in any number of commercial transactions and in cryptographic digital transactions as well as cryptographic protocols.

III. Proposed System

We explore which functionalities can be implemented in a way that participating parties have the incentive to provide their true private inputs upon engaging in the corresponding SMC protocols. This assumption is very general since most existing works in the area of privacy-preserving data Analysis assumes either all participating parties are honest (or semi-honest) or the majority of participating parties are honest. In our proposed system, we can prevent the participating parties from modifying their private inputs. In data analysis model, we use the Deterministic NCC (DNCC) technique in the proposed system. This DNCC technique does not use the Trusted Third Party (TTP). Instead of TTP, this DNCC technique checks the inputs given by the parties whether it is true or not on its own. Here also the input is sent in the form of query by the participating parties. But, the organization's name is not sent as in the case of existing system. The incentive to be checked is sent by the respective party. The DNCC will find the IP address of the system or party. It will then reach the party and many authorization steps will be carried by both the system and PPDA. Finally the inputs are checked and the response is given to the respective party. This reduces the computation time and this method is very easy to perform. Unlike the existing system, we can work without prior knowledge about the schema, structure, vocabulary and technical details of these sources.

A.User Interface Design

User Interface Design plays an important role for the user to move login window to user window. This module is created for the security purpose. User name and password willcheck username and password matches or not (valid username and valid password).

If we enter any wrong username or wrong password we can't enter into login window from user window. It generates some error message. So we are preventing from unauthorized user entering into the login window from user window. It will provide a good security for our project. Here the master node is the server.

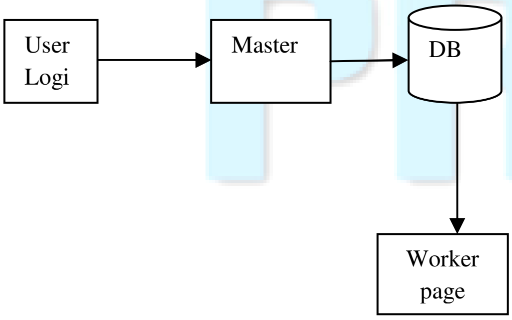

Fig. 1 Module diagram of User Interface Design

B. Create Number of Parties

We create n number of parties. Here *n* number of parties will send their inputs to single data analysis. The data analysis will store their inputs either horizontal or vertical partitions. Each party's information or incentives are stored in its own database and retrieved whenever required.

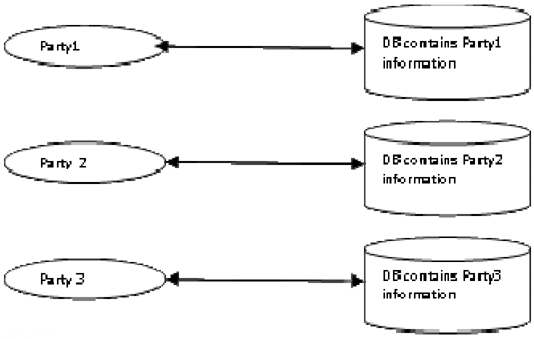
Party1
DB contains Party1 information
Party 2
DB contains Party2 information
Party 3
DB contains Party3 information

Fig. 2 Module to create number of parties

C. Data Analysis

Our Data Analysis is designed using cryptographic techniques. Data are generally assumed to be either vertically or horizontally partitioned. If parties choose horizontal partition then they input data for many different individuals. Same way if parties choose horizontal partition then they input data for different feature sets.

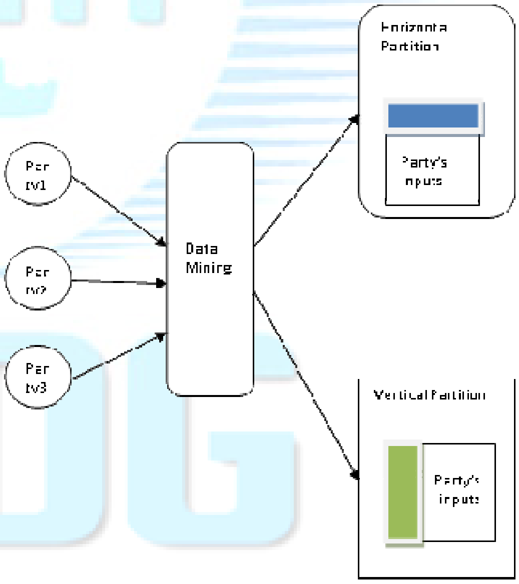

Fig. 3 Data Analysis Module

D. Input Computational Model

This model is designed for computing all the truthful inputs of all participating parties. Correctness: Here we have assumptions like the first priority for every participating party is to learn the correct result. Exclusiveness: Another one is, if possible, every participating party prefers to learn the correct result exclusively.

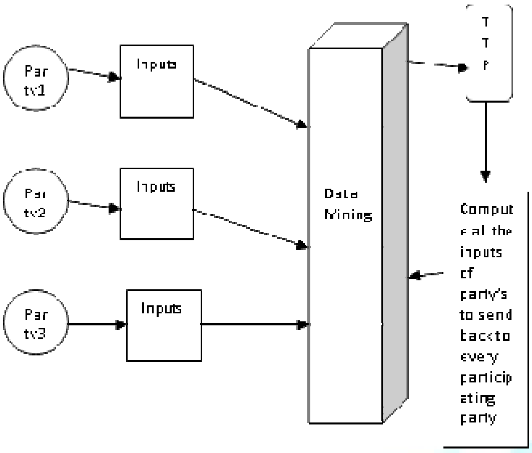
Party1
Inputs
Party2
Inputs
Party3
Inputs
Data Mining
TTP
Computes all the inputs of party's to send back to every participating party

Fig. 4 Module for Input Computational Model

E. Association Data Mining

It analyzes whether the association rule mining can be done in an incentive compatible manner. If it gets in the requested query then it search where it is located (either horizontal partition or vertical partition) Retrieve the result from partition. Result sends to particular party.

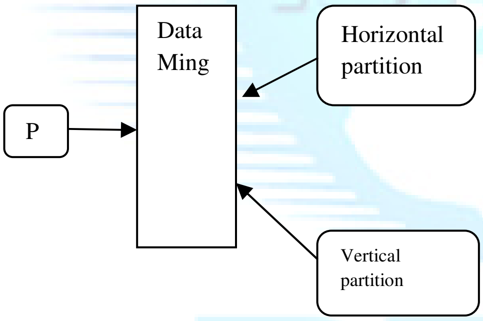

Fig. 5 Module for Association Data Mining

IV. System Model

The architecture of this system would explain the system description of this work. The system includes user login which is the first module, database to be accessed by the parties, *n* number of parties involved in computation, DNCC module, association rule mining module,data analysis module which contains two types of partitioning. 1)horizontal partitioning and 2)vertical partitioning.

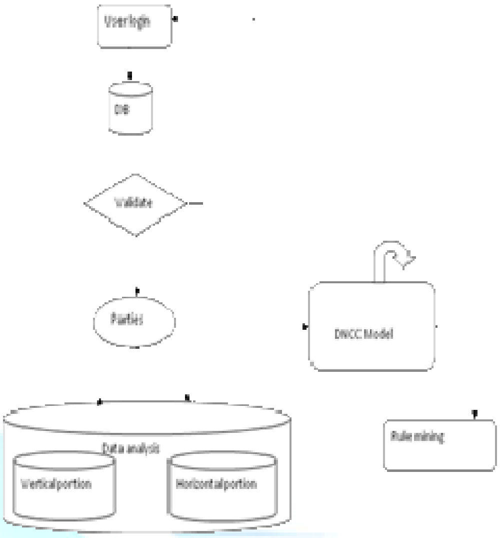

Fig. 6 Architecture of this system

A. Validate

The user log in by giving user id and password. If the user id and password is valid, user can access the database. Then the parties can collaborate share and store their data. Then the parties those are validated send their inputs or incentives to PPDA.

Here, the input is the worker name and the password. If it is valid, then the output for the validation module or the user interface design module is the worker window. This is the simple module where hackers can try to hack the password by all the possible methods. But in this paper, we are using NCC which maintains the cryptographic techniques. By using this techniques, security can be provided to the organization's information.

B. Creation of parties

Here, number of parties are created and the number depends upon the organizations that are interested in computation. They connect themselves to PPDA(single data analysis module). Then they perform computation on their private incentives to know the result correctly and exclusively. Here, the input and output is as follows.

The input is sending request to create a party. This request is sent to PPDA to join in computation. Then, second input is involving itself in communication and sharing information with other parties. The party sends request in the form of query by using its incentives.

The output is the creation of party with its own database. If any party do not wish to communicate, then it need not give request to PPDA.

C. Single data analysis module

The parties that are validated store their data in single data analysis module. Data analysis module someday will turn into a data warehouse since more inputs are sent by the parties. Here, the data is stored in either vertical or horizontal partition. The partition depends upon the type of data sent by the party. Then, this data or inputs are sent to DNCC module to check whether the information is true or not. In the input computation, all the party's inputs are computed. This acts as the input. Then, the output is to learn the correct result by the respective parties.

In the data analysis module, the input is the incentives sent by the parties to PPDA. The output is to store the data in either horizontal or vertical partitions.

D. Association rule mining

It's defined as number of data items in the data set. Let $I = \{i_1,i_2,….,i_n\}$ be a set of items. Let *DB* be a set of transactions, where each transaction *T* is an item set. Our data mining is summarize the association rule mining and analyze whether the association rule mining can be done in an incentive compatible manner over horizontally or vertically partitioned database.

Association rule mining is used to find whether the incentive is present in the query or not. It is a rule that is a collection of data set of the data items. Here, each data set contains number of data items and each data item contains number of attributes. We represent association rule mining as *I* (containing data sets of the data items). First, the item is checked whether it is present in the data set. Then, the incentives are checked whether they are present. Similarly, all the inputs are checked in *I*to find the association rule results.

Here the input is sending the request to data analysis module and reporting to the master. The master is the server or the home page. The worker nodes are the users. Then, the output is to retrieve the information and sending back the retrieved information to the respective organization or the party.

V. Conclusion and Future Work

Even though PPDA techniques guaranty that nothing other than the final result is revealed, whether or not the participating parties provide true inputs to PPDA cannot be verified. In this paper, we consider the parties to be semi-honest in sending their inputs to the single data analysis module. Therefore, for the parties that misbehaves, NCC technique is used. As this technique takes more computation time and also the computation cost is very high, we adopt DNCC technique that is easy to use and less time consuming. Also, DNCC knows the schema, structure beforehand. Trusted third parties are used in existing system along with NCC technique. But this gives rise to a question of how to provide security to the organization's information where the trusted third party is a dishonest one. Therefore, various authentication steps take place before accessing the incentives of an organization in DNCC. Finally, the security is provided for the incentives. Not only the security, confidentiality is also given by DNCC technique. A PPDA task can have many variations, and one common variation can be done at the last step for making PPDA method more secure. We investigate what kinds of PPDA tasks are incentive compatible under NCC model. Therefore, we extend the techniques and methods under NCC. We also aim to create more efficient secure multiparty computation techniques for implementing the data analysis tasks that are in DNCC so that the drawbacks of SMC technique is overcome. Therefore, the paper will need to consider the factor which discussed above in the future.

References

[1] J. Halpern and V. Teague, "Rational Secret Sharing and Multiparty Computation: Extended Abstract," Proc. Ann. ACM Symp.*Theory of Computing* (STOC '04), pp. 623-632, 2007.

[2] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," Proc.IEEE*Int'l Conf. Data Mining Workshop Privacy, Security, and Data Mining,* C. Clifton and V. Estivill-Castro, eds.,vol. 14, pp. 1-8, Dec. 2004.

[3] I. Ashlagi, A. Klinger, and M. Tenneholtz, "K-NCC: Stability Against Group Deviations in Non-Cooperative Computation,"*Proc. Third Int'l Conf. Internet and Network Economics,* pp. 564-569,2002.

[4] M. Kantarcioglu and O. Kardes, "Privacy-Preserving Data Mining in the Malicious Model," *Int'l J. Information and Computer Security,*vol. 2, pp. 353-375, Jan. 2001.

[5] G. Kol and M. Naor, “Cryptography and Game Theory: Designing Protocols for Exchanging Information,” *Proc. Conf. Theory of Cryptography*, p. 320, 2008.

[6] R. Agrawal and E. Terzi, “On Honesty in Sovereign Information Sharing,” *Proc. Int’l Conf. Advances in Database Technology,* pp. 240-256, 2006.

[7] S. Han and W.K. Ng, “Preemptive Measures against Malicious Party in Privacy-Preserving Data Mining,” *Proc. SIAM Int’l Conf. Data Mining (SDM),* pp. 375-386, 2008.