IMPLEMENTATION OF PRIVACY PRESERVING TECHNIQUE IN DISTRIBUTED DATA ANALYSIS IN DATA MINING

Srikanth Bethu
Department of Computer Science and Engineering
GRIET College, JNTU Hyderabad, India

R Aruna Flarence
Department of Computer Science and Engineering
GRIET College, JNTU Hyderabad, India

Suresh Mamidisetti
Department of Technical Education,
Government Polytechnic College, Hyderabad, Telangana, India

ABSTRACT

Information and Communication Technology has a challenging issues like Privacy and Security in maintaining confidentiality of a given data. The communicating and sharing of data has many benefits, and it provides great value to research in building accurate data analysis models. In this paper we discuss on sharing of data and participating parties which provide truthful inputs. The assumption of fact finding is based on the protocol that builds data analysis models which provides results from all participating parties. The participants who are not interested can be negotiated from producing results for analysis. In this paper credit card application is taken as example and analyzed deeply how distributed data analysis is taken place.

Key word: Data Analysis, Distributed Data, Protocols, Privacy and securit.

http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=9&IType=3

1. INTRODUCTION
Department of Energy supports research on building much more efficient diesel engines. Such an ambitious task requires the collaboration of geographically distributed industries, national laboratories and universities. Those institutions (including the potentially competing industry partners) need to share their private data for building data analysis [1] models that enable them to understand the underlying physical phenomena. Similarly, different pharmaceutical
companies may want to combine their private research data to predict the effectiveness of some protein families on certain diseases.

On the other hand, an omniscient data source eases misuse, such as the growing problem of identity theft. To prevent misuse of data, there is a recent surge in laws mandating protection of confidential data, such as the European Community privacy standards, U.S. health-care laws, and California SB1386. However, this protection comes with a real cost through both added security expenditure and penalties and costs associated with disclosure. For example, Card Systems was terminated by Visa and American Express after having credit card information stolen. Choice Point stock lost 20% of its value in the month following their disclosure of information theft. Such public relations costs can be enormous and could potentially kill a company. From lessons learned in practice, what we need is the ability to compute the desired “beneficial outcome” of data sharing [1] for analyzing without having to actually share or disclose data. This would maintain the security provided by separation of control while still obtaining the benefits of a global data source. Secure multi-party computation (SMC) has recently emerged as an answer to this problem. Informally, if a protocol meets the SMC definitions, the participating parties learn only the final result and whatever can be inferred from the final result and their own inputs. A simple example is Yao’s millionaire problem: two millionaires, Alice and Bob, want to learn who is richer without disclosing their actual wealth to each other. Recognizing this, the research community has developed many SMC protocols [2], for applications as diverse as forecasting, decision tree analysis and auctions among others. Nevertheless, the SMC model does not guarantee that data provided by participating parties are truthful. In many real life situations, data needed for building data analysis models are distributed among multiple parties with potentially conflicting interests. For instance, a credit card company that has a superior data analysis model for fighting credit card fraud may increase its profits as compared to its peers. An engine design company may want to exclusively learn the data analysis models that may enable it to build much more efficient diesel engines. An exclusive use of better data analysis models for predicting drug effectiveness may reduce the drug development time for a pharmaceutical company, which in return may translate into a huge competitive advantage. Clearly, as described above, building data analysis models is generally performed among parties that have conflicting interests. In the SMC model [3], we generally assume that participating parties provide truthful inputs. This assumption is usually justified by the fact that learning the correct data analysis models or results is in the best interest of all participating parties. Since SMC-based protocols require participating parties to perform expensive computations, if any party does not want to learn data models and analysis results, the party should not participate in the protocol. Still, this assumption does not guarantee the truthfulness of the private input data when participating parties want to learn the final result exclusively. For example, a drug company may lie about its private data so that it can exclusively learn the data analysis model. Although SMC protocols guarantee that nothing other than the final data analysis result is revealed, it is impossible to verify whether or not participating parties are truthful about their private input data. In other words, unless proper incentives are set, current SMC techniques cannot prevent input modification by participating parties. In order to better illustrate this problem, we consider a case from management where competing companies (e.g., Texas Instruments, IBM and Intel) establish a consortium (e.g., Semiconductor Manufacturing Technology1). The companies send the consortium their sales data, and key manufacturing costs and times. Then the consortium analyzes the data and statistically summarizes them in a report of industry trends, which is made available back to consortium members. In this case, it is in the interest of companies to learn true industry trends while revealing their private data as little as possible. Even though SMC protocols can prevent the revelation of the private data, they do not guarantee that companies send their true sales data and other required information.
1.1. Proposed Solution
In this paper, we analyze what types of distributed functionalities could be implemented in an incentive compatible fashion. In other words, 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. Each of these deals with the problem of ensuring truthfulness in data mining. However, each one requires the ability to verify the data after the calculation. Although verification based techniques are very useful, there are cases where verification is not feasible due to legal, social and privacy concerns.

We first develop key theorems, then based on these theorems, we analyze what types of privacy-preserving data analysis tasks could be conducted in a way that telling the truth is the best choice for any participating party. Secure multi-party computation [4] (SMC) has recently emerged as an answer to this problem. Our main motive is Avoid the fraud entry.

There are two modules namely user and admin. The user can enter the website and create their own account for free of cost they can perform transaction and purchase the items they require. The full control over the users and their functions and the main function of the admin is to unblock the blocked users.

![Figure 1 Conceptual Diagram](image)

The user can have access to the website only by logging into the website by using a key that is generated to the user by using random key generation algorithm. The key will be sent to the user to the provided mail-id at the time of registering. If the key entered is correct then the user can perform further transactions or the user will be asked to enter the key again. If the user enters the wrong key for more than three times then the user will be blocked and the blocked user cannot have access to the website unless the admin unblocks the use.

First computation node will start running. After party node enter user name and password that is validated by compatible node. Then computation node assigns the work to the data mining nodes. Data mining [5] node finishes his work and reposted to the compatible node. TTP collects the inputs of parties and group of parties input for particular work presented by party nodes. The motto is to provide privacy by invoking incentives.
2. RELATED WORK
In this section, we begin with an overview of Privacy Preserving Distributed Data Analysis. Then we briefly discuss the concept of Non-Cooperative Computation. Table 1 Provides Common Notations and Terminologies used extensively for the rest of this paper. In Addition, the terms Secure And Privacy-Preserving are interchangeable thereafter.

2.1. Privacy Preserving Data Analysis
Many privacy-preserving [6] data analysis [6] protocols have been designed using cryptographic techniques [7]. Data are generally assumed to be either vertically or horizontally partitioned. In the case of horizontally partitioned data, different sites collect the same set of information about different entities. For example, different credit card companies may collect credit card transactions of different individuals. Privacy-preserving distributed protocols have been developed for horizontally partitioned data for building decision trees, mining association rules, and generate k-means clusters [9] and k-nn classifiers. In the case of vertically partitioned data, we assume that different sites collect information about the same set of entities, but they collect different feature sets. For example, both a university pay roll and the university’s student health center may collect information about a student. Again, privacy-preserving protocols for the vertically partitioned case have been developed for mining association rules, building decision trees and k-means clusters. To the best of our knowledge, all the previous privacy preserving data analysis protocols assume that participating parties are truthful about their private input data. Recently, game theoretical techniques have been used to force parties to submit their true inputs. The techniques developed in assume that each party has an internal device that can verify whether they are telling the truth or not. In our work, we do not assume the existence of such a device. Instead, we try to make sure that providing the true input is the best choice for a participating party.

2.2. Literature Survey
Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy of company strength. Once these things are satisfied, ten next steps are to determine which operating system and
language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above considerations are taken into account for developing the proposed system. Input design consists of development and specifications and procedures for data preparation, the steps necessary to put transaction data into a usable form for processing and data entry, the activity of data into the computer processing. The five objectives of input design are:

1. Controlling the amount of input
2. Avoiding delay
3. Avoiding error in data
4. Avoiding extra steps
5. Keeping the process simple

3. DATA MODEL DESIGNING ISSUES AND ANALYSIS

Following are the modules that were developed for designing architecture. They are as follows:

- User Interface Design
- Create Multiple Organizations
- Data Analysis and Integration
- Inputs computation model
- Association Data Mining

3.1. User Interface Design

In this module we create a user page using Graphical User Interface (GUI), which will be the media to Connect user with the server and through which client can able to give request to the server and server can send the response to the client, through this module we can establish the communication between client and server using webpage. A program interface that takes advantage of the computer’s graphics capabilities to make the program easier to use. Well-designed graphical user interfaces can free the user from learning complex command languages. Well-designed graphical user interfaces can free the user from learning complex command languages. On the other hand, many users find that they work more effectively with a command-driven interface, especially if they already know the command language.

Its goal is to enhance the efficiency and ease of use for the underlying logical design of a stored program. Thus the user interacts with information by manipulating visual widgets that allow for interactions appropriate to the kind of data they hold. The widgets of a well-designed interface are selected to support the actions necessary to achieve the goals of the user.

3.2. Create Multiple Organizations

Here we are design no. of parties. Each and every party may have information to store their database. All the parties may send their inputs to Data Analysis [8] module. Here all n no. of parties will send their inputs to single data analysis. The data analysis will store their inputs either horizontal or vertical partitions. In this module we can create no. of parties. Each and every party may have own database it can store their information either vertical portion or horizontal portion.
3.3. Data Analysis and Integration
Our Data Analysis designed using cryptographic techniques. Data are generally assumed to be either vertically or horizontally partitioned. In the case of horizontally partitioned data, different sites collect the same set of information about different entities. In the case of vertically partitioned data, we assume that different sites collect information about the same set of entities. A party can store their input data either vertical partition or horizontal partitioned. If parties choose horizontal partition then the input data for many different individuals. Same way if parties choose horizontal partition then the input data for many different individuals.

3.4. Inputs Computation model
This model to design for compute all the truthful inputs of all participating parties here going to assumptions like the first priority for every participating party is to learn the correct result. Another one is, if possible, every participating party prefers to learn the correct result exclusively.

3.5. Association Data Mining
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. If get in the requested query then it search where it is located either horizontal partition or vertical partition retrieve the result from partition after that result send to particular party.

3.6. Test Procedure
The procedure for testing this screen is planned in such a way that the data entry, status calculation functionality, saving and quitting operations are tested in terms of GUI testing, Positive testing, Negative testing using the corresponding GUI test cases, Positive test cases, Negative test cases respectively.

Example: In the registration form, if we are giving characters in the text field of Contact No then we get an error as data truncated at the column by name Contact No. why because, Contact No only having the digits but not the characters. So to avoid this problem, when the creation of database we are given the data type of Contact No as „integer‟. This is tested here.

4. RESULTS ANALYSIS
In the above home page fig. 3. The user can register, login, contact us, and know about us. The user can know about us and the services provided from the information provided in the about us page. Here the user can contact us from the information provided in the contact us page.

**Figure 4** Register page

In the above register page fig. 4 the user has to register by giving user details like name, uname, date of birth, address, contact number, security question, security answer, email id. In this the user gives his mail id and the password is sent to the given user mail id mainly for the security purpose.

**Figure 5** Login page

In the above login page fig. 5 The user has to login to the website by entering the user name and the password which is sent to the user mail id.
In the above user home page fig.6 the user can know about his/her personal details like the amount in his account, his personal details like which were provided during the time of registration.

In the above fund transfer page fig. 7 the user can transfer the amount to another user at this particular page by entering the user name and amount to whom to be sent and should also give correct security question and answer to perform transactions. If in case the given security question or answer is wrong then he needs to enter it again. If this takes place for more than 3 times then the user gets blocked and he cannot have access to the account.
In the above purchase page fig.8 the user can purchase the product he/she wishes by clicking on the product where the details will be provided. To purchase the product the user needs to enter the security question and answer. If in case the given security question or answer is wrong then he needs to enter it again. If this takes place for more than 3 times then the user gets blocked and he cannot have access to the account. In purchase page the user can purchase either the electronics or home needs. The electronics and home needs pages are as follows: In the above electronics page 4.6 the user can select the product by clicking on the product and purchase the product.

In the above view transactions page Fig.9 the user can view all the transactions performed along with the date and time.
5. CONCLUSION AND FUTURE SCOPE

In this paper, we analyzed the types of distributed functionalities implemented in an incentive compatible fashion. In other words, we explored the functionalities that 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.

Even though privacy-preserving data analysis techniques guarantee that nothing other than the final result is disclosed, whether or not participating parties provide truthful input data cannot be verified. In this paper, we have investigated what kinds of PPDA tasks is incentive compatible under the NCC model. Based on our findings, there are several important PPDA tasks that are incentive driven. Users have to give their truth full data for security system. User only knows the answers for security questions. None Other than the user can know about the security questions and answers provided by the user. User’s knows the Fraud entry. If fraud user try to access the user account then he cannot perform any transactions as the fraud user does not know the security questions the fraud user will be blocked and the user cannot have access to the account unless and until the user is unblocked by the admin. Fraud could be detected.

REFERENCES


