

# Privacy Preserving Data Analysis Technique

G.Monika<sup>1</sup>, R.Saraswathi<sup>2</sup>, K.Sujitha<sup>3</sup>, Mrs.M.Varalakshmi<sup>4</sup>Student, Department of CSE, Manakula Vinayagar Institute of Technology, Puducherry, India<sup>1,2,3</sup>Assistant Professor, Department of CSE, Manakula Vinayagar Institute of Technology, Puducherry, India<sup>4</sup>

**Abstract--** In many cases, competing parties who have private data may collaboratively conduct Fraud detection tasks to learn beneficial data models or analysis results. For example, different credit card companies may try to build better models for credit card fraud detection through Fraud detection tasks. Similarly, competing companies in the same industry may try to combine their sales data to build models that may predict the future sales. In many of these cases, the competing parties have different incentives. Although certain fraud detection techniques guarantee that nothing other than the final 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, even current Fraud detection techniques cannot prevent participating parties from modifying their private inputs. This raises the question of how to design incentive compatible Fraud detection techniques that motivate participating parties to provide truthful input data. In this paper, we first develop key theorems, then base on these theorem, we analyze what types of Fraud detection tasks could be conducted in a way that telling the truth is the best choice for any participating party.

## I. INTRODUCTION

Privacy and security, particularly maintaining confidentiality of data, have become a challenging issue with advances in information and communication technology. The ability to communicate and share data has many benefits, and the idea of an omniscient data source carries great value to research and building accurate data analysis models. For example, for credit card companies to build more comprehensive and accurate fraud detection system, credit card transaction data from various companies may be needed to generate better data analysis models. 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 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. 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 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, for applications as diverse as forecasting, decision tree analysis and auctions among others. Nevertheless, the SMC model does not guarantee that data provided by

## International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization,

Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)  
On 10<sup>th</sup> & 11<sup>th</sup> February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHANICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

participating parties are truthful. In many real life situations, data needed for building data 2 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 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

### II. RELATED WORK & BACKGROUND

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 I provides common notations and terminologies used extensively for the rest of this paper. In addition, the terms *secure* and *privacy-preserving* are interchangeable thereafter.

#### A. Privacy-Preserving Data Analysis

Many privacy-preserving data analysis protocols have been designed using cryptographic techniques. Data are generally assumed to be either vertically or horizontally partitioned. (Table II shows a trivial example of different data partitioning schemes.) 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 and k-n n classifiers. (See for a survey of the recent results.) 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 . (See for a survey of the

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.

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, we recent results.) 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.

#### B. Non-Cooperative Computation

Recently, research issues at the intersection of computer science and game theory have been studied extensively. Among those research issues, algorithmic mechanism design and non-cooperative computation are closely related to our work. The field of algorithmic mechanism design tries to explore how private preferences of many parties could be combined to find a global and socially optimal solution[20]. Usually in algorithmic mechanism design, there exists a function that needs to be maximized based on the private inputs of the parties, and the goal is To devise mechanisms and payment schemes that force individuals to tell their true private values. In our case, since it is hard to measure the monetary value of the data analysis results, devising a payment scheme that is required by many mechanism design models is not viable that is designed for parties who want to jointly compute the correct function results on their private inputs. Since data analysis algorithms can be seen as a special case, modifying non-cooperative computation model for our purposes is a natural choice .The non-

## International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization,

Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)  
On 10<sup>th</sup> & 11<sup>th</sup> February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHANICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

cooperative computation (NCC) model can be seen as an example of applying game theoretical ideas in the distributed computation setting. In the NCC model, each party participates in a protocol to learn the output of some given function  $f$  over the joint inputs of the parties. First, all participating parties send their private inputs securely to a trusted third party (TTP), then TTP computes  $f$  and sends back the result to every participating party. The NCC model makes the following assumptions:

1) **Correctness**: the first priority for every participating party is to learn the *correct* result;

### III. EXISTING WORK

Forecasting is increasingly being applied to business decision making. Many forecasting methods (for example, see ) have been developed, such as time-series techniques and regression techniques. Collaborative forecasting allows different entities to jointly perform business forecasting where each entity contributes its own data. As pointed out in , collaborative forecasting, in comparison to traditional forecasting, gives better productivity and portability throughout the supply chain. Collaborative forecasting has been extensively studied by many companies organizations, and academia. Most of the solutions either assume existence of a central planner who has all the information about the system, or assume that each participant of the computation shares

### IV. CONCLUSION:

In this work, we provided privacy-preserving solutions to collaborative forecasting and benchmarking that can be used to increase the reliability of local forecasts and data correlations, and to conduct the evaluation of local performance compared to global trends. We gave both building blocks and their use in protocols for a number of different forecasting methods based on time-series and regression techniques. The building blocks are general enough to be used in other protocols for forecasting and benchmarking, as well as in other applications. In particular, the division protocols presented in this work, to the best of our knowledge, are the first attempt to perform division in secure multi-party

2) **Exclusiveness**: if possible, every participating party prefers to learn the *correct* result *exclusively*. In other words, learning the correct result is the most important objective of every party. Other factors such as privacy and voyeurism could be also considered in the NCC setting. We omit such discussion here. Additional details can be found in [18]. In this paper, we use the NCC setting where each party wants to learn the data mining result *correctly*, if possible prefers to learn it *exclusively*. Also, we assume that revealing only the result does not violate privacy.

all of her information with other participants. These solutions, however, are problematic when the data is sensitive and the participants are reluctant to share their private, proprietary information. Our approach is to perform collaborative forecasting in a privacy-preserving manner, therefore eliminates the above concern. The problem of secure forecasting and benchmarking is closely related to secure multi-party computation . The SMC problem was introduced by Yao and extended by Goldreich, Micali, Wigderson and others to list a few). Goldreich states in that although the general secure multi-party computation problem is solvable in theory, using the solutions derived by these general results for special cases can be impractical. In other words, efficiency dictates development of special solutions for special cases.

computation as well as to perform computations on floating point numbers. This work can be extended in a number of ways. Future directions include:

- The model can be extended to other time-series forecasting techniques.
- Along with providing short-range forecasting, we would like to be able to perform long-range forecasts. Long-range forecasts take into account seasonal changes and other long-range patterns.
- We also would like to design protocols to cover other types of regressions for benchmarking collaboration. This will allow us to draw reliable conclusions for different types of data distributions.
- We would like to make some of the protocols provided in this paper more robust against other types of malicious behavior.

**International Journal of Innovative Research in Science, Engineering and Technology**

An ISO 3297: 2007 Certified Organization,

Volume 3, Special Issue 1, February 2014

**International Conference on Engineering Technology and Science-(ICETS'14)  
On 10<sup>th</sup> & 11<sup>th</sup> February Organized by**

Department of CIVIL, CSE, ECE, EEE, MECHANICAL Engg. and S&amp;H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

## REFERENCES:

- [1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *VLDB '94*, pages 487–499, Santiago, Chile, September 12–15 1994. VLDB.
- [2] Rakesh Agrawal and Evimaria Terzi. On honesty in sovereign information sharing. In *EDBT*, pages 240–256, 2006.
- [3] Mikhail J. Atallah, Marina Bykova, Jiangtao Li, and Mercan Karahan. Private collaborative forecasting and benchmarking. In *Proc. 2d. ACM Workshop on Privacy in the Electronic Society (WPES)*, Washington, DC, October 28 2004.
- [4] B. Chor and E. Kushilevitz. A zero-one law for boolean privacy. In *STOC '89*, pages 62–72, New York, NY, USA, 1989. ACM Press.
- [5] www.doe.gov, doe news, feb. 16 2005.
- [6] Wenliang Du and Zhijun Zhan. Building decision tree classifier on private data. In Chris Clifton and Vladimir Estivill-Castro, editors, *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, pages 1–8, Maebashi City, Japan, December 9 2002. Australian Computer Society.
- [7] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, No L(281):31–50, October 24 1995.
- [8] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, 1990.
- [9] O. Goldreich, S. Micali, and A. Wigderson. How to play a mental game - a completeness theorem for protocols with honest majority. In *19th ACM Symposium on the Theory of Computing*, pages 218–229, 1987.
- [10] Oded Goldreich. *The Foundations of Cryptography*, volume 2, chapter General Cryptographic Protocols. Cambridge University Press, 2004.
- [11] Joseph Halpern and Vanessa Teague. Rational secret sharing and multiparty computation: extended abstract. In *STOC '04*, pages 623–632, New York, NY, USA, 2004. ACM Press.
- [12] Standard for privacy of individually identifiable health information. *Federal Register*, 67(157):53181–53273, August 14 2002.
- [13] Geetha Jagannathan and Rebecca N. Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In *Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593–599, Chicago, IL, August 21–24 2005.
- [14] Murat Kantarcioglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE TKDE*, 16(9):1026–1037, September 2004.
- [15] Xiaodong Lin, Chris Clifton, and Michael Zhu. Privacy-preserving clustering with distributed EM mixture modeling. *Knowledge and Information Systems*, 8(1):68–81, July 2005.
- [16] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Advances in Cryptology – CRYPTO 2000*, pages 36–54. Springer-Verlag, August 20–24 2000.
- [17] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3):177–206, 2002.
- [18] Robert McGrew, Ryan Porter, and Yoav Shoham. Towards a general theory of non-cooperative computation (extended abstract). In *TARK IX*, 2003.
- [19] Moni Naor, Benny Pinkas, and R. Sumner. Privacy preserving auctions and mechanism design. In *Proceedings of the 1st ACM Conference on Electronic Commerce*. ACM Press, 1999.
- [20] Noam Nisan and Amir Ronen. Algorithmic mechanism design (extended abstract). In *STOC '99*, pages 129–140, New York, NY, USA, 1999. ACM Press.
- [21] Yoav Shoham and Moshe Tennenholtz. Non-cooperative computation: boolean functions with correctness and exclusivity. *Theor. Comput. Sci.*, 343(1-2):97–113, 2005.
- [22] Jaideep Vaidya and Chris Clifton. Privacy preserving association rule mining in vertically partitioned data. In *ACMSIGKDD'02*, pages 639–644, Edmonton, Alberta, Canada, July 23–26 2002.
- [23] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Rec.*, 33(1):50–57, 2004.
- [24] Andrew C. Yao. Protocols for secure computation. In *Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science*, pages 160–164. IEEE, 1982.
- [25] Andrew C. Yao. How to generate and exchange secrets. In *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, pages 162–167. IEEE, 1986.