ISSN 2229-5518

Privacy-Preserving Data Analysis using Incentive Compatability

S.Sabiya Begum, E.Naga Bhavani ,T.Sireesha, S.Shafi, J.David Sukeerthi Kumar

Abstract—In many cases, competing parties who have private data may collaboratively conduct privacy preserving distributed data analysis (PPDA) 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 PPDA 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 PPDA 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 PPDA techniques cannot prevent participating parties from modifying their private inputs. This raises the question of how to design incentive compatible privacy-preserving data analysis 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 privacy-preserving data analysis tasks could be conducted in a way that telling the truth is the best choice for any participating party.

Index Terms—Privacy, Secure multi-party computation, Non-cooperative computation.

1.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 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.

2.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, 16], mining association rules, [14], and generate k-means clusters [15] and k-nn classifiers. (See [23] 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, [22], building decision trees [6] and k means clusters [13]. (See [23] for a survey of the 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

IJSER

theoretical techniques have been used to force parties to submit their true inputs [2]. The techniques developed in [2] 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.

2.1Non-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 (e.g., Vickrey-Groves-Clarke mechanisms [20]).

Instead, we adopt the non-cooperative computation model [21] 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-cooperative compu ation (NCC) model can be seen as an example of applying game theoretical ideas in the distributed computation setting [21]. 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;
2) ***Exclusiveness***: if possible, every participating party prefers to learn the *correct* result *exclusively*.

Under the *correctness* and *exclusiveness* assumptions, the NCC model is formally defined as follows: Given a set of n parties, for a party i, we denote its private input as vi ∈ Di, where Di is the domain of the possible inputs of party i. For simplicity, we assume that all Di = D for all i. Parties joint input is represented as v = (v1, . . . , vn), where v ∈ Dn.We use v−i to represent (v1, . . . , vi−1, vi+1, . . . vn), and (vi, v−i) to denote the reconstruction of v. It is also assumed that the v values are distributed according to some probability function, and the probability of seeing any v ∈ Dn is always nonzero.

In the NCC model, for calculating any n party function f : Dn 7−→ R with range R, we use the following simple protocol:

1) The TTP computes f(v′) = f(v′1 , . . . , v′ n) and sends the results back to the participating parties;
2) Each party i computes f(v) based on f(v′) received from TTP and vi. Considering the above simple protocol does not limit its generality. Under the literature of SMC, the TTP can be replaced such that the required functionality (represented by f) is still computable without violating privacy regarding each participating party's private input [11]. The next definition states the conditions a function needs to satisfy under the NCC model.

4.ANALYZING DATA ANALYSIS TASKS IN THE NCC MODEL

So far, we have developed techniques to prove whether or not a function is in DNCC. Combining the two concepts DNCC and SMC, we can analyze privacy preserving data analysis tasks (without utilizing a TTP) that are incentive compatible. We next prove several IEEE Transactions on Knowledge and Data Engineering, (Volume:25 , Issue: 6)June 2013 7 such important tasks (as function with Boolean output, set operations, linear functions, etc) that either satisfy or do not satisfy the DNCC model. Also, note that the data analysis tasks analyzed next have practical SMC implementations.

3.Function with Boolean Output

From SMC literature, we know that there are few functions that can be evaluated if the adversary controls −1 parties. Here, we prove that functions with Boolean outputs that are n − 1 private are not in DNCC.

3.1Theorem 1:

A function from f : D1×D2×· · ·× Dn 7→ {0, 1} is n−1-private if there exits a protocol f so that no coalition of size ≤ n−1 can infer any additional information from the execution, other than the function result. Further more, f is n − 1 private if and only if it can be represented as:

$$f(v1, v2, \ldots, vn) = f1(v1) \oplus f2(v2) \oplus \cdots \oplus fn(vn)$$

where fis are arbitrary functions with boolean outputs and ⊕ is the binary XOR operation.

3.2Theorem 2:

There does not exit any non-constant n − 1 private function with boolean output that is in *DNCC*.

Proof: According to Theorem.1, we know that any n − 1 private function is of the form:
f(v1, v2, . . . , vn) = f1(v1) ⊕ f2(v2) ⊕ · · · ⊕ fn(vn). Clearly, for any ti, we can define the gi as:

$$g_i(f(t_i(S_i), S_{-i}), S_i) = f(t_i(S_i), S_{-i}) \cup S'$$
$$= (\cup_{j=i} S_j) \cup (S_i \setminus S') \cup S'$$
$$= S_1 \cup \cdots \cup S_n = f(S_i, S_{-i})$$

5.CONCLUSION

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 are incentive compatible under the NCC model. Based on our findings, there are several important PPDA tasks that are incentive driven. As a future work, we will investigate incentive issues in other data analysis tasks, and extend the proposed theorems under the probabilistic NCC model. The PPDA tasks analyzed in the paper can be reduced to evaluation of a single function. Now, the question is how to analyze whether a PPDA task is in DNCC if it is reduced to a set of functions. In other words, is the composition of a set of DNCC functions still in DNCC? We will formally answer this question in the future. Another important direction that we would like to pursue is to create more efficient SMC techniques tailored towards implementing the data analysis tasks that are in DNCC.

6.FUTURE WORK

Even though privacy-preserving data analysis tech-niques 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 are incentive compatible under the NCC model. Based on our findings, there are several important PPDA tasks that are incentive driven. As a future work, we will investigate incentive issues in other data analysis tasks, and extend the proposed theorems under the probabilistic NCC model.

The PPDA tasks analyzed in the paper can be reduced to evaluation of a single function. Now, the question is how to analyze whether a PPDA task is in DNCC if it is reduced to a set of functions. In other words, is the composition of a set of DNCC functions still in DNCC? We will formally answer this question in the future.

Another important direction that we would like to pur-sue is to create more efficient SMC techniques tailored towards implementing the data analysis tasks that are in DNCC.

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 clas-sifier on private data. In Chris Clifton and Vladimir

IJSER

Estivil l-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. Australia n 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 I.(281):31–50, October 24 1995.

[8] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recog-nition*. Academic Press, San Diego, CA, 1990.

[9] O. Goldreich, S. Micali, and A. Wigderson. How to play any 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 Univer-sity 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 heal th informa-tion. *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 Con-ference on Knowledge Discovery and Data Mining*, pages 593– 599, Chicago, IL, August 21-24 2005.

[14] Murat Kantarcıo˘glu and Chris Clifton. Privacy-prese rving dis-tributed 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 com-putation: boolean functions with correctness and exclusivity. *Theor. Comput. Sci.*, 343(1-2):97–113, 2005.

[22] Jaideep Vaidya and Chris Clifton. Privacy preserving asso-ciation rule mining in vertically partitioned data. In *ACM SIGKDD'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 *Proceed-ings 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.

IJSER