Incentive Compatible Privacy-preserving Distributed Data Sharing and Mining

Murat Kantarcioğlu
University of Texas at Dallas, Department of Computer Science

1 Introduction

Privacy and security, particularly maintaining confidentiality of data, have become a challenging issue with advances in information and communication technology. In addition, the ability to communicate and share data has many benefits as well as obvious security risks. For example, Department of Energy supports research on building much more efficient diesel engines [1], which would require the collaboration of geographically distributed industries, national laboratories, and universities. Those institutions (including the competing industrial partners with potentially conflicting incentives) need to share their private data for building data mining models that enable them to understand the underlying physical phenomena.

Secure multi-party computation (SMC) techniques [2] have recently emerged as one of the answers to secure distributed data sharing and mining. 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. Nevertheless, the SMC model does not guarantee that data provided by participating parties are truthful. Even more, due to the security guarantees provided by the SMC protocols, a party can lie about its data without even getting caught. For example, Alice can easily inflate her net worth to one trillion dollars and due to the security provided by the SMC protocols, Bob will only be able to learn that Alice is richer.

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 mining 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 mining 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. These examples indicate that there are many cases where distributed data sharing needs to be performed among parties that have conflicting interests.

In the SMC model, 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 or results. Although SMC protocols guarantee that nothing other than the final data mining 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 Technology [1]). The companies send the consortium their sales data, and key manufacturing costs and times. In turn, 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 disclosure of the private data, it does not guarantee that companies send their true sales data and other required information. To see this consider the following example.

1 www.sematech.org
Example 1.1 Assume that $n$ companies would like to learn the total sales for a particular type of product. Let $x_i$ be the $i^{th}$ company’s sales amount. In order to estimate the total sales for the overall industry, companies need to calculate $s = \sum_{i=1}^{n} x_i$. Any company may exclusively learn the correct result by lying about its input. Company $i$ may report $x'_i$ instead of the correct $x_i$. Given the wrong $s'$ (computed based on $x'_i$ and truthful values from the other parties), the company $i$ can calculate the correct sum $s$ by setting:

$$s = s' + x_i - x'_i$$

As illustrated above, any company may have the incentive to lie about its input in order to learn the result exclusively, and at the same time, the correct result (e.g., $s$) can be computed from its original input, modified input and the incorrect final result (e.g., $x_i$, $x'_i$ and $s'$). If this situation always occurred, no company would have the incentive to be truthful. Fortunately, the complexity of a function that is computed determines whether or not the situation (demonstrated by the above example) could occur. To further illustrate the importance of incentives in data sharing consider the following example.

Example 1.2 Assume that $n$ individuals who want to ride on a private plane need to securely calculate the sum of their weights \footnote{For the sake of example, assume that weight information is sensitive.} to figure out how much fuel is needed for the plane. Clearly, if a person lies about his/her weight, it may endanger everyone on the plane. Therefore, any rational person who will ride on the plane will have an incentive to tell the truth.

Please note that in both cases described above, we need to calculate sum of certain values (e.g. sum of sales vs sum of weights). In the case of competing companies, due to incentive structure, no one has an incentive to tell the truth. In the case of individuals who are riding on a plane, everyone has an incentive to tell the truth. These two examples illustrate the importance of the incentive structure in determining the quality of the information that individuals and their organizations choose to share.

Given the various incentives, the goal should be to design incentive compatible privacy-preserving data sharing protocols that incorporate the incentives of the participating parties to ensure privacy-preserving, efficient and truthful protocols. For this attempt to succeed, we need to first develop techniques to estimate the final result of the data sharing and data mining results so that each participant can in advance know whether it makes sense to participate in the data sharing protocols. Second, given the utilities of each participant, we need to analyze whether the underlying task can be achieved in an incentive compatible manner. For example, in the above discussion, we argued that summation cannot be done in an incentive compatible manner if every participant wants to learn the result exclusively. To address this challenge, we may use techniques from computational game theory (e.g. non-cooperative computation \cite{Shoham&Tennenholtz}) to analyze incentive issues in various distributed data sharing tasks.

Finally, by considering the incentive structure of the participants, we need to develop new and efficient secure distributed data sharing and data mining algorithms. To tackle this issue, we may use various techniques such as monetary punishments to cheaters, repeated game theoretical models and so on to create secure, efficient and truthful distributed data sharing and mining protocols.

In summary, we believe that incentive issues are as critical as security and privacy in data sharing and new techniques are need to address such incentive issues.

References

