1 Introduction

Privacy and security, particularly maintaining confidentiality of data, have become challenging issues 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 more accurate data analysis models. For example, various hospitals may need to share their privacy sensitive genomic data along with the hospital discharge data for building data mining models that enable them to understand the underlying relationship between genes and diseases. Similarly, different pharmaceutical companies may want to combine their privacy sensitive research data to predict the effectiveness of certain protein families on some diseases.

Secure multi-party computation (SMC) techniques [7] have recently emerged as one of the answers to privacy-preserving distributed data mining and data integration problems. For example, privacy-preserving SMC protocols are provided for distributed data integration [5], distributed data mining [15], and distributed schema matching [19]. Recognizing the potential of SMC tools, the research community has developed many other SMC protocols, for applications as diverse as forecasting [4], electronic commerce and auctions [17] among others.

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 [21]: two millionaires, Alice and Bob, want to learn who is richer without disclosing their actual wealth to each other.) These privacy protection guarantees are provided by employing cryptographic tools that involve expensive operations such as modular exponentiation [11]. Although, the SMC based methods allow trade-offs between efficiency and privacy protection by letting users to choose the size of the cryptographic keys (i.e., using public key encryption schemes with key size 512 bits is much faster than with public key size 1024 bits), there is no general and easy way for trade-offs between accuracy and privacy. In addition, in many cases, SMC based approaches do not scale well for large amounts of data [11].

On the other hand, sanitization/anonymization based techniques allow organizations to reveal privacy-sensitive data under some privacy guarantees by distorting the data (i.e., using privacy metrics that measure the amount of privacy protection). For example, Census Bureaus reveal sanitized versions of private demographic information. Over the years, a plethora of methods such as noise addition [3, 12], k-anonymity [18], l-diversity [16], t-closeness [14], m-invariance [20], and differential privacy [6] have been proposed to perturb and sanitize data to protect individual privacy. In these methods, higher levels of protection typically translate into further deviation from the original data and, consequently, less accurate results. Usually these sanitization techniques provide parameters (e.g., k in k-anonymity and ε in differential privacy) to trade off between accuracy and privacy. Although sanitization techniques are quite efficient, they could require significant amount of data distortion to preserve privacy.

To illustrate the problems with the current approaches, consider the case of a medical researcher who wants to integrate patient data sets from two different hospitals to compute correlation between a certain gene and smoking. Currently, such a researcher may try to use some of the existing SMC techniques to integrate such data sets. To our knowledge, all the existing pure SMC based techniques (e.g., [2, 13, 1]) would require $O(n \cdot m)$ cryptographic operations where n (resp. m) is the number of records in the first (resp. second) data set. If $n = m = 10000$, such an integration task will require $10^8$ cryptographic operations. Even with the help of cryptographic accelerators such a protocol will take a few weeks to complete [10]. As an alternative approach, a researcher could try to use sanitized data sets that could be released by the hospitals (in real life, under HIPAA rules [8], such data sets could be released for research purposes after removing certain identifiers.) to do the integration. Clearly, using such sanitized data sets directly will not give accurate results. Instead, in this position paper, we advocate an integrated approach. It is possible that by looking at the sanitized data sets, we may know that certain records will not match (i.e., if one record has an age value 50 and the other record has 10, we can conclude that they could not belong to the same person) and use SMC techniques (i.e., secure euclidean distance calculation) for the record pairs that could potentially match. Our initial results [9] indicate that such an integrated approach could reduce the total running time to 1% of the pure SMC techniques without any reduction in accuracy.

---

1 We use sanitization and anonymization interchangeably.
Clearly, by using such an integrated approach, our ultimate goal is to preserve the individual privacy (e.g., protecting the patients’ privacy) instead of preventing the disclosure of any sensitive information related to the data holder (e.g., protecting the trade secrets of the hospitals). As the above example indicates, an integrated approach is orthogonal to any new sanitization and SMC techniques. If new and more efficient SMC techniques are developed for some particular task (e.g., secure euclidean distance calculation), we will be able to use it to give even more efficient solutions using such an integrated framework. In addition, this approach could be applied to other domains such as homeland security and intelligence applications where large amounts of distributed data need to be analyzed in a privacy-preserving way under some resource constraints. (i.e., privacy-preserving join of airline passenger lists and terrorist watch lists [9].)

To create an integrated framework for privacy-preserving data analytics that leverages existing sanitization and SMC techniques to create efficient solutions for large data sets requires addressing major research challenges. First, we need to create new definitions and techniques such that privacy guarantees provided by the developed algorithms can be analyzed. Second, we need to provide additional information (e.g., some useful statistics about the unperturbed data) together with the sanitized data to facilitate approximate computation of the required results without violating the original privacy definitions. Finally, we need to develop more efficient privacy-preserving protocols that leverage the existence of the sanitized data to maximize accuracy under given resource constraints.

Overall, we believe that such an integrated approach will accelerate the use of privacy-preserving data analytics tools by enabling low cost privacy-preserving solutions.

References


