ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (85-89), Month: April - June 2015, Available at: www.researchpublish.com

A Unique Approach for Multiparty Data Distribution Using Anonymization Protocol

[1]Kiruthika Murugesan, [2]Saranya Sivasamy, [3]Sudha Elangovan

[1, 2, 3] B. Tech- Information Technology (Student- Final year), V.S.B Engineering College, Karur, India

Abstract: **Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information Data integration between autonomous entities should be conducted in a way that no more information than necessary entity is revealed between the participating entities. New knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration. To securely integrate person-specific sensitive data from multi data providers, the integrated data still retain the essential information for supporting data mining tasks. The multi-party data publishing generate an integrated data table satisfying differential privacy. The algorithm also satisfies the security definition in the secure multiparty computation.**

Keywords: **Differential Privacy, Secure Data Integration, Classification Analysis.**

1. INTRODUCTION

Large amount of databases exist today due to the rapid advances in communication and storing systems. Each database is owned by a particular autonomous entity, for example, medical data by hospitals, income data by tax agencies, financial data by banks, and census data by statistical agencies. Moreover, the emergence of new paradigms such as cloud computing increases the amount of data distributed between multiple entities. These distributed data can be integrated to enable better data analysis for making better decisions and providing high-quality services. For example, data can be integrated to improve medical research, customer service, or homeland security. However, data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time, new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration. In this paper, we propose an algorithm to securely integrate person-specific sensitive data from multiple data providers, where by the integrated data still retain the essential information for supporting data mining tasks.

The following real-life scenario further illustrates the need for following real-life scenario further illustrates the need for simultaneous data sharing and privacy preservation person-specific sensitive data. This research problem was discovered in a collaborative project with the financial industry. We generalize their problem as follows: A bank A, financial company B and a loan company C have different sets of attributes about the same set of individuals identified by the common identifier attribute (ID), such that bank A owns DA (ID; Job; Balance),while financial company B owns DB (ID;Account No;) while loan company C owns DC(ID; Sex; Salary). These parties want to integrate their data to support better decision making such as loan or credit limit approvals. In addition to parties A, B and C their partnered credit card company D also has access to the integrated data, so all three parties A, B, C and D are data recipients of the final integrated data. Parties A, B and C have two concerns. First, simply joining DA, DB and DC would reveal sensitive information to the other party. Second, even if DA, DB and DC individually do not contain person-specific or sensitive information, the integrated data can increase the possibility of identifying the record of an individual.

According to the taxonomy presented in Fig. 1 (ignore the dotted line for now) so that this individual becomes one of many female professionals. However, Machanavajjhala et al. [34] have pointed out that with additional knowledge about the victim, k-anonymous data are vulnerable to background knowledge attacks. To prevent such attacks, ‘-diversity requires that every QID group should contain at least ‘“well-represented” values for the sensitive attribute. Similarly, there are a number of other partition-based privacy models such as ð_; kÞanonymity [56], ðc; kÞ-safety [35], and t-closeness [32] that differently model the adversary and have different assumptions about her background knowledge. However, recent research has indicated that these privacy models are vulnerable to various privacy attacks [54], [60], [19], [28] and provide insufficient privacy protection.

2. RELATED WORKS

Data privacy has been an active research topic in the statistics, database, and security communities for the last three decades [17]. The proposed methods can be roughly categorized according to two main scenarios:-Interactive versus non interactive.

In an interactive framework, a data miner can pose queries through a private mechanism, and a database owner answers these queries in response. In a non interactive framework, a database owner first anonymizes the raw data and then releases the anonymized version for data analysis. Once the data are published, the data owner has no further control over the published data. This approach is also known as privacy preserving data publishing (PPDP). Single versus multiparty. Data may be owned by a single party or by multiple parties. In the distributed (multiparty) scenario, data owners want to achieve the same tasks as single parties on their integrated data without sharing their data with others. Our proposed algorithm addresses the distributed and non interactive scenario.

Single-party scenario: We have already discussed different privacy models in Section 1. Here, we provide an overview of some relevant anonymization algorithms. Many algorithms have been proposed to preserve privacy, but only a few have considered the goal for classification analysis. Iyengar has presented the anonymity problem for classification and proposed a genetic algorithmic solution. Bayardo and Agrawal have also addressed the classification problem using the same classification metric of. Fung et al. [18] have proposed a top-down specialization (TDS) approach to generalize a data table. LeFevre et al. have proposed another anonymization technique for classification using multidimensional recoding. More discussion about the partition-based approach can be found in the survey of Fung et al. Differential privacy has recently received considerable attention as a substitute for partition-based privacy models for PPDP. However, so far most of the research on differential privacy concentrates on the interactive setting with the goal of reducing the magnitude of the added noise [11], [14], [47], releasing certain data mining results [4], [8], [9], [16], or determining the feasibility and infeasibility results of differentially-private mechanisms [5], [53], [36]. Research proposals [2], [23], [38], [58] that address the problem of non interactive data release only consider the single-party scenario.

Therefore, these techniques do not satisfy the privacy requirement of our data integration application for the financial industry. A general overview of various research works on differential privacy can be found in the survey of Dwork Distributed interactive approach. This approach is also referred to as privacy preserving distributed data mining (PPDDM) [10]. In PPDDM, multiple data owners want to compute a function based on their inputs without sharing their data with others. This function can be as simple as a count query or as complex as a data mining task such as classification, clustering, and so on.

For example, multiple hospitals may want to build a data mining model for predicting disease based on patients’ medical history without sharing their data with each other. In recent years, different protocols have been proposed for different data mining tasks including association rule mining [50], clustering [51], and classification [33], [6]. However, none of these methods provide any privacy guarantee on the computed output (i.e., classifier, association rules). On the other hand, Dwork et al. [13], and Narayan and Haeberlen [43] have proposed interactive algorithms to compute differentially private count queries from both horizontally and vertically partitioned data, respectively distributed non interactive approach. This approach allows anonymizing data from different sources for data release without exposing the sensitive information. Jurczyk and Xiong [27] have proposed an algorithm to securely integrate horizontally partitioned data from multiple data owners without disclosing data from one party to another. Mohammed et al. [41] have proposed a distributed algorithm to integrate horizontally partitioned high dimensional health care data. Unlike the distributed anonymization problem for vertically partitioned data studied in this paper, these methods [27], [41] propose algorithms for horizontally partitioned data.

3. PRIVACY MODEL

Differential privacy is a recent privacy definition that provides a strong privacy guarantee. It guarantees that an adversary learns nothing more about an individual, regardless of whether her record is present or absent in the data.

Differential Privacy: [14]. A randomized algorithm Ag is differentially private if for all data sets D and D0, where their symmetric difference contains at most one record.

4. SECURITY MODEL

In this section, we briefly present the security definition in the semi honest adversary model. Additionally, we introduce the required cryptographic primitives that are instrumented inside the proposed algorithm in this paper

4.1 Secure Multiparty Computation:

Security with respect to semi honest behaviour): Two probability distributions are computationally indistinguishable if no efficient algorithm can tell them apart. Namely, the output distribution of every efficient algorithm is oblivious whether the input is taken from the first distribution or from the second distribution [20]. Many of the protocols, as in the case of the proposed algorithm in this paper, involve the composition of secure sub protocols in which all intermediate outputs from one sub protocol are inputs to the next sub protocol. These intermediate outputs are either simulated given the final output and the local input for each party or computed as random shares. Random shares are meaningless information by themselves. However, shares can be combined to reconstruct the result. Using the composition theorem [21], it can be shown that if each sub protocol is secure, then the resulting composition is also secure.

4.2 Cryptographic Primitives:

Yao's Protocol [59]. It is a constant-round protocol for secure computation of any probabilistic polynomial-time function in the semi honest model. Let us assume that we have two parties, P1 and P2, with their inputs x and y, respectively. Both parties want to compute the value of the function fðx; yÞ. Then, P1 needs to send P2 an encrypted circuit computing fðx; :Þ. The received circuit is encrypted and accordingly P2 learns nothing from this step. Afterwards, P2 computes the output fðx; yÞ by decrypting the circuit. This can be achieved by having P2 obtaining a series of keys corresponding to its input y from P1 such that the function fðx; yÞ can be computed given these keys and the encrypted circuit. However, P2 must obtain these keys from P1 without revealing any information about y. This is done by using the oblivious transfer protocol [21].

5. MULTI -PARTY PROTOCOL FOR EXPONENTIAL MECHANISM

In this section, we present a Multi-party protocol for the exponential mechanism together with a detailed analysis. As discussed in Section 3, the exponential mechanism chooses a candidate that is close to optimum with respect to a utility function while preserving differential privacy. In the distributed setting, the candidates are owned by two parties and, therefore, a secure mechanism is required to compute the same output while ensuring that no extra information is leaked to any party.

5.1 Distributed Exponential Mechanism (DistExp)"

The distributed exponential mechanism presented inAlgorithm1 takes the following items as input: Finite discrete alternatives hðv1; u1Þ; . . . ; ðvn; unÞi, where a pair ðvi; uiÞ is composed of the candidate vi and its score ui. Parties P1 and P2 own ðv1; u1Þ; . . . ; ðvj; ujÞ and ðvjþ1; ujþ1Þ . . . ðvn; unÞ, respectively. . Privacy budget.

6. MULTI-PARTY DIFFERENTIALLY PRIVATE DATA RELEASE ALGORITHM

In this section, we first define some notations, state the problem, and present our assumptions. We then describe the two-party algorithm for differentially private data release for vertically partitioned data.

6.1 Preliminaries:

Suppose two parties P1 and P2 own data table D1 and D2, respectively. Both parties want to release an integrated anonymous data table ^D(Apr 1 ; . . .;Apr d ;Acls) to the public for classification analysis. The attributes in D1 and D2 are classified into three categories: 1) An explicit identifier attribute Ai that explicitly identifies an individual, such as SSN and Name. These attributes are removed before releasing the data. 2) A class attribute Acls that contains the class value, and the goal of the data miner is to build a classifier to accurately predict the value of this attribute. 3) A set of predictor attributes Apr ¼ fApr 1 ; . . .;Aprd g, whose values are used to predict the class attribute. The explicit identifier and the class attribute are shared among the two parties. Given a table D1 owned by P1, a table D2 owned by P2 and a privacy parameter _, our objective is to generate an integrated anonymous data table ^D such that 1) ^D satisfies _-differential privacy and 2) the algorithm to generate ^D satisfies the security definition of the semi honest adversary model.

7. COST ESTIMATE

Most of the computation and the communication take place during the execution of DistExp (Line 7) and SSPP (Line 20). The runtime of the other steps is less than 30 seconds for Adult data set. Hence, we only elaborate the runtime of DistExp and SSPP.

7.1 Distributed Exponential Mechanism:

As discussed in Section 5, the computation and the communication complexity of the distributed exponential mechanism are dominated by the cost of the comparison circuit. In the following, we provide an estimate for the computation and the communication costs of evaluating the COMPARISON circuit. Here, we assume that P1 encodes and P2 evaluates the encrypted circuit. The roles of P1 and P2 can be swapped.

7.2 Secure Scalar Product Protocol:

We adopt the Secure Scalar Product Protocol of [26] and use its reported running time to estimate the cost of this step for our algorithm. The primary cost of SSPP depends on the number of homomorphic encryptions that is equal to jDj, the size of the data set. As reported in [26], the estimated cost of the homomorphic encryptions is 19.5 s on average when jDj ¼ 30162 (the size of our data set) on Intel Xeon 3-GHz processor.

8. CONCLUSION

In this paper, we have presented the first two-party differentially private data release algorithm for vertically partitioned data. We have shown that the proposed algorithm is differentially private and secure under the security definition of the semi honest adversary model. Moreover, we have experimentally evaluated the data utility for classification analysis. The proposed algorithm can effectively retain essential information for classification analysis. It provides similar data utility compared to the recently proposed single party algorithm [38] and better data utility than the distributed k-anonymity algorithm for classification analysis.

9. FUTURE ENHANCEMENTS

The proposed algorithm is only applicable for the two-party scenario because the distributed exponential algorithm and the other primitives (e.g., random value protocol, secure scalar product protocol) are limited to a two-party scenario. The proposed algorithm can be extended for more than two parties by modifying all the sub protocols while keeping the general top-down structure of the algorithm.

Semi honest Adversary Model. This is the common security definition used in the SMC literature [26]; it is realistic in our problem scenario because different organizations are collaborating to securely share their data for mutual benefits. However, they may be curious to learn additional information from the messages they received during the protocol execution. To extend the algorithm for malicious parties, all sub protocols should be extended and must be secure under the malicious adversary model.

REFERENCES

[1] N. Mohammed, B.C.M. Fung, P.C.K. Hung, and C. Lee,“Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service,” Proc. ACM Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’09), 2009.

[2] M. Naor and B. Pinkas, “Efficient Oblivious Transfer Protocol,” Proc. 12th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA ’01), 2001.

[3] .A. Narayan and A. Haeberlen, “DJoin: Differentially Private Join Queries over Distributed Databases,” Proc. 10th USENIX Conf. Operating Systems Design and Implementation (OSDI’12), 2012.

[4] P. Paillier, “Public-Key Cryptosystems Based on Composite Degree Residuosity Classes,” Proc. 17th Int’l Conf. Theory and Application Cryptographic Techniques, pp. 223-238, 1999.

[5] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[6] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, pp. 557-570, 2002.

[7] .A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,“‘-Diversity: Privacy Beyond k-Anonymity,” ACM Trans. Knowledge Discovery from Data, vol. 1, article 3, 2007.

[8] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern, “Worst-Case Background Knowledge in Privacy- Preserving Data Publishing,” Proc. IEEE Int’l Conf. Data Eng. (ICDE ’07), 2007