Security on Two-Party Differentially Private Data

C.S Venkatraja[1] S.Nagaraj[2]
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]Selvam College of Technology, Namakkal, Tamilnadu, India.

Abstract— In this project study the problem of answering queries about private data that is spread across multiple different databases. For instance, a medical researcher may want to study a possible correlation between travel patterns and certain types of illnesses. The necessary information exists today suppose in airline reservation systems and hospital records but it is maintained by two separate companies who are prevented by law from sharing this information with each other, or with a third party. This separation prevents the processing of such queries, even if the final answer, e.g., a correlation coefficient, would be safe to release. Here present DJoin, a system that can process such distributed queries and can give strong differential privacy guarantees on the result. DJoin can support many SQL queries, including joins of databases maintained by different entities. This experimental evaluation shows that DJoin can process realistic queries at practical timescales. This project can be viewed as a special case of a more general approach for producing synthetic data.

Key words: Natural Fiber Composites, Mechanical Properties

I. INTRODUCTION

Secure multiparty computation (smc) protocols are one of the first techniques used in privacy preserving data mining in distributed environments. The idea behind these protocols is based on the theoretical proof that two or more parties, both having their own private data, can collaborate to calculate any function on the union of their data. While doing so, the protocol does not reveal anything other than the output of the function or anything that can be computed from it in polynomial time. Moreover, the protocol does not require a trusted third party. While these properties are promising for privacy preserving applications, SMC may be prohibitively expensive

In fact, many SMC protocols for privacy preserving data mining suffer from high computation and communication costs. Furthermore, those that are closest to be practical are designed for the semi honest model, which assumes that parties will not deviate from the protocol. Theoretically, it is possible to convert protocols in the semi honest model into protocols in the malicious model. However, the resulting protocols are even more costly.

The major contributions of this project are a privacy preserving association rule mining algorithm given a privacy preserving scalar product protocol, and protocol for computing scalar product while preserving privacy of the individual values. I show that it is possible to achieve good individual security with communication cost comparable to that required to build a centralized data warehouse. There are several directions for future research. Handling multiple parties is a non-trivial extension, especially if we consider collusion between parties as well. This work is limited to boolean association rule mining. Non-categorical attributes and quantitative association rule mining are significantly more complex problems.

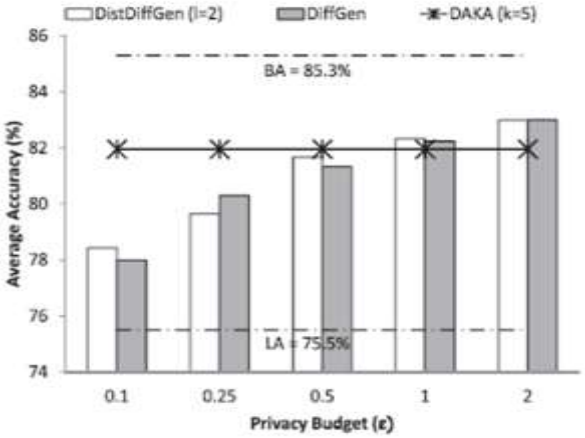

Fig. 1: Privacy Budget

II. RELATED WORK

Data de-identification reconciles the demand for release of data for research purposes and the demand for privacy from individuals. This paper proposes and evaluates an optimization algorithm for the powerful de-identification procedure known as k-anonymization. A k-anonymized dataset has the property that each record is indistinguishable from at least k-1 others. Even simple restrictions of optimized -anonymity are NP-hard, leading to significant computational challenges. We present a new approach to exploring the space of possible anonymizations that tames the combinatory of the problem, and develop data-management strategies to reduce reliance on expensive operations such as sorting.

III. EXISTING SYSTEMS

This research problem was discovered in a collaborative project with the financial industry. We generalize their problem as follows: A bank A and a loan company B have different sets of attributes about the same set of individuals identified by the common identifier attribute (ID), such that bank A owns DAðID; Job; Balance, while loan company Bowns DBðID; 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 and B, their partnered credit card company C also has access to the integrated data, so all three parties A, B, and C are data recipients of the final integrated data. Parties A and B have two concerns. First, simply joining DA and DB would reveal sensitive information to the other party. Second, even if DA and DB individually do not contain person-specific or sensitive information, the integrated data can increase the possibility of identifying the record of an individual.

IV. PROPOSED WORK

In this paper, we propose an algorithm to securely integrate person-specific sensitive data from two data providers, whereby the integrated data still retain the essential information for supporting data mining tasks. The following real-life scenario further illustrates the need for simultaneous data sharing and privacy preservation of person-specific sensitive data. To prevent such linking attacks in the existing system we have proposed algorithms that enable

A. Modules description

1) Secure Multiparty Computation Design

Retrieving the contents of a multi-valued attribute from a group, such as a distribution list, can often result in a large number of returned values. Although the data can be encrypted to disallow unauthorized access, encryption does not prevent the service provider from monitoring the I/O activities of user queries, thus inferring (and potentially misusing) sensitive information of corporate or personal importance. We aim to support efficient database querying in such an environment, while offering a high degree of protection for the privacy of user queries from the database server.

They propose a mechanism to construct a secure encryption scheme to enforce the SCs and show that finding an optimal, secure encryption scheme for a given set of SCs is NP-hard They propose a specific scheme for the metadata to be maintained by the server, to support efficient query processing.

It consists of two key components:

- The structural metadata called the discontinuous structural interval index (DSI) that pertains to the structure of the database;
- a B-tree index as the metadata on the values that is based on transforming the unencrypted values by splitting and scaling techniques, so that the distribution of unencrypted data is different from that of ciphered data model requires that an individual should not be identifiable from a group of size smaller than k based on the quasi-identifier (QID), where QID is a set of attributes that may serve as an identifier in the data set. two parties to integrate their data satisfying the k-anonymity privacy model.

B. K- Anonymity Formulation

K-Anonymity is a well-known privacy preservation technique proposed in to prevent linking attacks on shared databases. Linking attacks are performed by adversaries who know some attributes of an individual to identify him/ In the data set. A database is said to be k-anonymous if every tuple projected over the quasi-identifier attributes appears at least k times in the data base. k-Anonymization is the process of enforcing the k-anonymity property on a given database by using generalization and suppression of values. They complement our analytical results with a battery of experiments anonymization "benchmark" that we are aware of. This dataset was prepared as described by Iyengar to the best of our ability. It consists of 9 attributes (8 regular and one class column) and 30,162 records, and contains actual census data that has not already been anonymized.

C. Generalization and Suppression

Merging the similar data types of a given selected mining algorithm into a generalized data type seems to be a good approach to reduce the transformation complexity in SMC Protocols. The Generalization process, including merging and transforming phases is proposed. In the merging phase, the original data types of the data sources to be mined are first merged into the generalized one. The transforming phase is then used to convert the generalized data types into the target ones for the selected mining algorithm. Some of the data are not visible to the parties to improve the performance of k-anonimity using suppression methodology.

D. Evaluation

One goal of our experiments is to understand the performance of K-OPTIMIZE. Beyond this, the ability to quickly identify optimal solutions allows exploring many other interesting aspects of k-anonymization. The dataset used in our experiments was the adult census dataset from the Irvine machine learning repository,since this dataset is the closest to a common k-granularity and query size and shape on query processing efficiency, as well as the client post-processing cost, demonstrating the effectiveness of our approach. This paper studies the problem of protecting the key scope of range queries that are executed on untrusted database servers.

1) Formal Statement of Our Contribution

Our contribution, as suggested by the paper's title, comes in the parts of privacy, accuracy, and consistency, each of which are critical components of any data analysis system. At an intuitive level, which we soon formalize, we are concerned with

- Privacy: The presence or absence of any one data element should not substantially influence the distribution over outcomes of the computation.
- Accuracy: The difference between the reported marginal and true marginal should be bounded, preferably independent of the size of the data set.
- Consistency: There should exist a contingency table whose marginal equal the reported marginal.

2) Comparison with Other Definitions.

Differential Privacy provides much stronger guarantees than other privacy definitions of which we are aware. For example, k- Nonetheless, neither of these definitions protects against even simple background knowledge of the form "My colleague Mr. R., who works in zip code 2770*, is in the database".

a) Accuracy

Privacy guarantees are of course meaningless without accompanying accuracy guarantees. We could easily erase the data if the former were all we cared about. We now detail guarantees that our algorithm makes about the accuracy of the counts in the released marginals, while ensuring differential privacy.

b) Consistency

The matter of consistency among the released marginals might appear trivial; indeed most previous approaches, which produced actual randomized data sets, it is a non-

issue, as their tables are produced from these specific data sets. However, there is previous work, namely, that assures differential privacy and strong accuracy simply by adding noise to related cell values. It is unlikely that there exists a single data set that yields all of the released marginals, and this potential inconsistency in the released data can be the source of many technical frustrations.

THEOREM 2. *[20]. For any $f : \mathcal{D} \to R^d$, the addition of Laplace noise[3] with variance $2\sigma^2$ preserves $(\Delta f/\sigma)$-differential privacy.*

PROOF. Using the definition of the Laplace density, the density at any a is

$$\mu[a|D] \propto \exp(\|f(D) - a\|_1/\sigma) \quad (4)$$

Applying the triangle inequality, we bound the ratio

$$\frac{\mu[a|D_1]}{\mu[a|D_2]} = \frac{\exp(\|f(D_1) - a\|_1/\sigma)}{\exp(\|f(D_2) - a\|_1/\sigma)} \quad (5)$$

$$\leq \exp(\|f(D_1) - f(D_2)\|_1/\sigma) . \quad (6)$$

The last term is bounded by $\exp(\Delta f/\sigma)$, by the definition of Δf. Thus (1) holds for singleton sets $S = \{a\}$, and the theorem follows by integrating over S. □

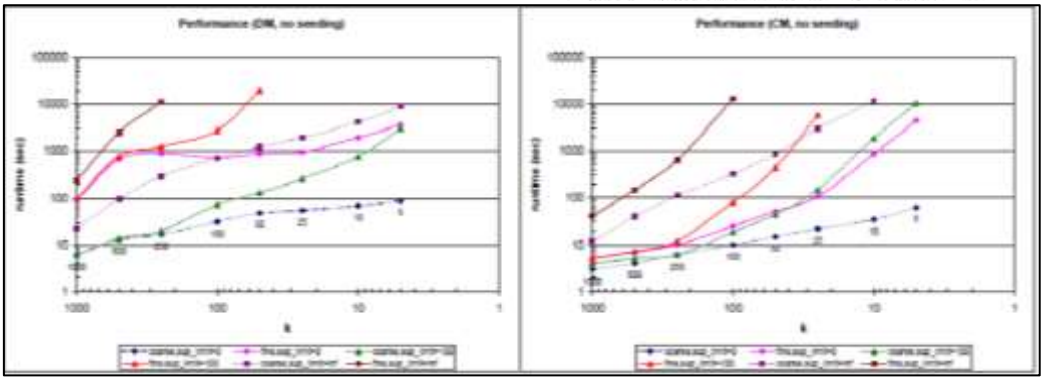

Fig. 6(a): Performance of the K-OPTIMIZE Procedure

V. TABLE PARTITIONING

Vertical partitioning involves creating tables with fewer columns and using additional tables to store the remaining columns. A common form of vertical partitioning is to split dynamic data (slow to find) from static data in a table where the dynamic data is not used as often as the static.

The database security is achieved using the vertical portioning by this module. If the adversary wants to hack the data, they have to compromise all the databases which are not easier process to perform the database injection attacks. The goal of this section is to provide a formal framework for constructing and evaluating algorithms and systems that release information such that the released information limits what can be revealed about properties of the entities that are to be protected.

However, the sites must not reveal individual transaction data. We present a two-party algorithm for discovering frequent item sets with minimum support levels, without either site revealing individual transaction values. Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as clustering, classification, association rule mining and sequence detection.

Vertical partitioning involves creating tables with fewer columns and using additional tables to store the remaining columns. A common form of vertical partitioning is to split dynamic data (slow to find) from static data (fast to find) in a table where the dynamic data is not used as often as the static. The database security is achieved using the vertical portioning by this module. If the adversary wants to hack the data, they have to compromise all the databases which are not easier process to perform the database injection attacks.

VI. CONCLUSION

Most SMC protocols are expensive in both communication and computation. We introduced a look-ahead approach for SMC protocols that helps involved parties to decide whether the protocol will meet the expectations before initiating it. We presented a look-ahead protocol specifically for the distributed k-anonymity by approximating the probability that the output of the SMC will be more utilized than their local anonymizations. Experiments on real data showed the effectiveness of the approach.

REFERENCES

[1] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing Across Private Databases," Proc. ACM Int'l Conf. Management of Data, 2003.

[2] C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. Advances in Data¬base Technology - EDBT-2004, 9th Int'l Conf. on Extending Database Technology, 183-199.

[3] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K.Talwar, "Privacy Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release," Proc. ACM Symp. Principles of Database Systems (PODS '07), 2007.

[4] R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. IEEE Int'l Conf. Data Eng. (ICDE '05), 2005.

[5] C. Dwork, "Differential privacy: A survey of results," in Proc. of the 5th Annual Conference on Theory and Applications of Models of computation (TAMC'08), Xi'an, China, LNCS, vol. 4978. Springer-Verlag, December 2008, pp. 1–19.

[6] U. M. Fayyad and K. B. Irani. Multi-interval discretiza¬tion of continuous-valued attributes for

classification learning. In Proc. of 13th Int'l Joint Conf. on Artificial Intelligence, 1022-1027, 1993.
[7] A. Hundpool and L. Willenborg. Mu-Argus and Tau Argus: Software for Statistical Disclosure Control. Third Int'l Seminar on Statistical Confidentiality, 1996.
[8] L. Kissner and D. Song. Privacy-preserving set operations. In Proc. CRYPTO, Aug. 05.
[9] 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.
[10] B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newsl., 4(2):12–19, Dec. 2002.
[11] 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.

