Data Publishing For Two-Party Sensitive Data Using Random Value Protocol

Dr.J.Jagadessan[1], D. Anow Fanny[2]
[1]*Prof., & Head of Department of CSE,* [2]*M.Tech(CSE) Scholar*
Department of Computer Science and Engineering, SRM University, Ramapuram Campus, Chennai, Tamil Nadu, India.

Abstract - Privacy of data is a difficult task to implement in the real world with full security. The data publishing of privacy-preserving describes the problem of disclosing sensitive data when mining for useful information. Even though many models exist in the real world for preserving the private data differential privacy provides the strongest privacy guarantees. The paper addresses the issues related to private data publishing, where there is a difference in the attributes for the same set of individuals are held by two parties. The algorithm defines the private release of vertically partitioned data between two parties in the context of semi-honest adversary model. The paper also presents a two-party protocol used for the exponential mechanism. The exponential mechanism is defined based on the secure multi-party computation algorithm that releases differentially private data in a secure way. Experimental results on real-life data suggest that the proposed algorithm can effectively preserve information in a secured manner and can be implemented for a data mining task.

Keywords: Differential Privacy, Secure Multi-party Computation, Two-Phase Validation.

I. INTRODUCTION

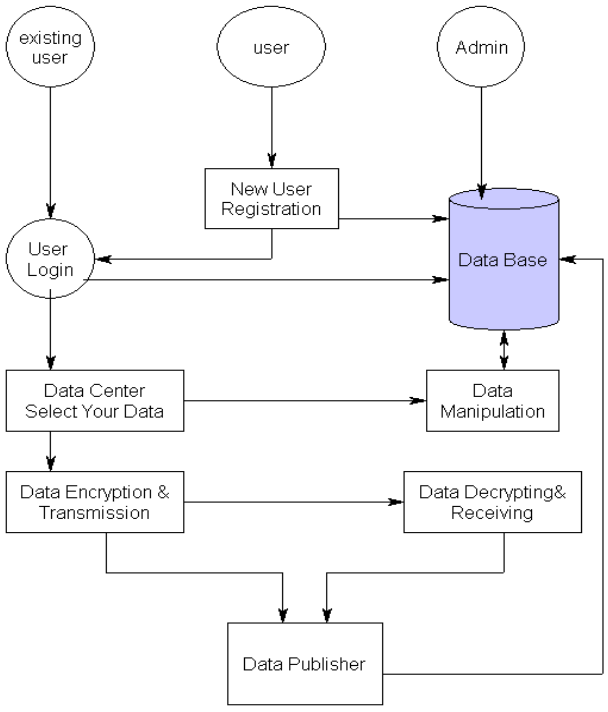
Many real-time applications involve computations that require sensitive data from two or more individuals to be integrated and published. For example, consider a genetic application with a genome database. Enormous information about the person-specific sensitive data will be contained in all genome databases that need full security. When a comparison is made between the genomes of two different persons are studied, it is necessary to preserve one's individual data from others. In case of providing treatment for a particular disease the data obtained from the data publisher of both parties must be most effective. The comparisons results in obtaining significant values, but are irrelevant because of the privacy concerns of both individual and study participants. The need that arises out of this is to produce the result of the comparison without exposing either party's private inputs. Outcome of the process is to make privacy-preserving computationally practical and accessible enough to be used in all real-time applications.

Secure multi-party computation and privacy-preserving are much related with each other. Persisting problems that arise in this context are (i) whether the computation of the function can be done safely and (ii) how to compute the results by minimizing the chances in leaking of original data. For example, it is always possible to collect all the data in one place and run an algorithm on the pooled data and then compute the results. A technique used in privacy-preserving data mining is under the study for a long time in the field of cryptography called secure multiparty computation (SMC). This problem deals with a setting where a set of parties with private inputs wishes to jointly compute some function of their inputs. Clearly, a protocol must be defined that provides the guarantee to solve privacy-preserving data mining problems.

The aim of secure multiparty computation is to enable parties to carry out such distributed computing tasks in a secure manner of machine crashes and other external faults. SMC is concerned with the possibility of deliberately malicious behavior by some external entity. In this context, it is assumed that, the outcome of execution of a protocol can be altered by an external entity or by the participating parties. The attack may happen to learn some private information or cause the result of the computation to be incorrect. In order to avoid the problems, privacy and correctness requirements factors of SMC is established to maintain the security. SMC assumes that a semi-honest adversary model is defined the specified protocol to provide better security between the parties. Another goal is to make secure computation more accessible to developers by developing programming tools for defining secure computations at a high level, based on flow of information analysis and program partitioning.

II. SYSTEM MODEL

Very large databases exist today due to the rapid advances in communication and storing systems. Data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. An algorithm is proposed for securing the integrated person-specific sensitive data from two data providers, but still retaining the essential information for supporting data mining tasks. The algorithm also satisfies the security definition in the secure multiparty computation literature. The problem statement defines the following: (i) Two –party protocol is used for the exponential mechanism where the first two-party data publishing algorithm for vertically partitioned data set is generated.(ii) Differential privacy provides a provable privacy guarantee that makes no assumption but proposes a top-down specialization(tds) approach to generalize a data table. In this paper we describe an implementation of the two-party case, using Random value protocol, and present various algorithmic protocol improvements. These optimizations are analyzed theoretically using experiments of various semi-honest adversarial situations. The protocol can be used as a sub protocol by any other algorithm that requires the exponential mechanism in a distributed setting.



The system provides authentication to both parties and the administrator. When a new user wants to access the service they must have to register themselves to obtain rights so that the system provides better security. Data manipulation means to deal with the information and try to find different patterns and to compare the values according to different criteria. This performs the major role actually more number of user have been using our service they are willing to store and manipulate their data in secured way. Sensitive data is everywhere. Organizations are taking a data-centric approach for the protection and control of their sensitive information, by preserving the overall efficiencies and economies of scale. Data center is used to store the bulk of data, Data centers store, manage, process, and contains all kinds of digital data and information, provides application services or management for various data processing such as web hosting, intranet, telecommunication and information technology. Two party data integration is the most important need in this system

III. PREVIOUS WORK

Now a day's more number of database is available because of rapid advances in communication and storing systems. Each database is owned by different parties and 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. Theoretical solutions to this problem have been defined long back by Andrew Yao based on garbled circuits. The garbled circuits of Yao's approach have traditionally been considered more of a theoretical curiosity than a practical mechanism for building privacy-preserving applications. Recent developments in the field of cryptographic techniques and newer implementation approaches are on the verge of change, however we need to admit the possibility of scalable and practical secure computation. For example, data can be integrated to improve medical research, customer service, or homeland security. 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.

Disadvantages:

- Difficult to securely integrate person-specific sensitive data from two data providers.
- Words, linking attack is possible while integrating.
- When integrating the two party data, there is the chance to loss the data.

- In existing system there is the chance for leaking of sensitive information or disclosing some important information.
- Here data utility is less and provides insufficient privacy protection.

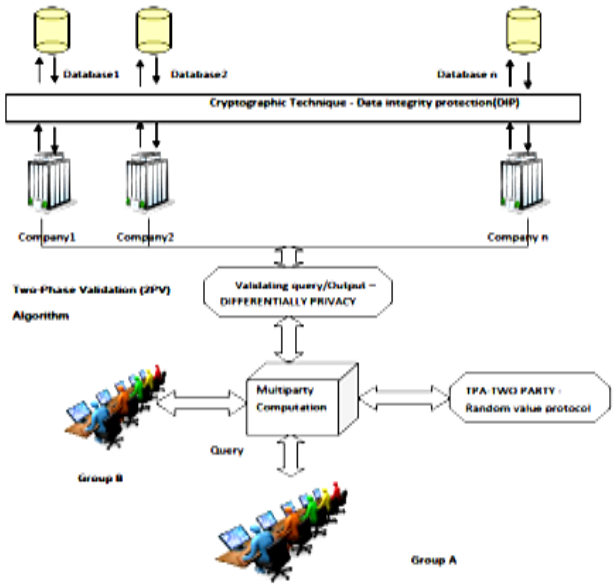
IV. PROPOSED METHODOLOGY

We deal with the problem of private data publishing, in which various attributes for the same set of individuals are held by two parties. The algorithm is defined for differentially private data of a vertically partitioned data between two parties in the semi-honest adversary model. To achieve this, we need to present a two-party protocol for the exponential mechanism. This protocol can act as a sub protocol by any other algorithm that requires the exponential mechanism in a distributed setting. More over in our proposed system a random value protocol is generated for the integration of the two party data. The results of experimental data suggest that the proposed algorithm can effectively preserve information for a data mining task. The paper also proposes privacy model that provides a provable privacy guarantee. The proposed algorithm can effectively retain essential information for classification analysis. It provides similar data utility when compared to other recently proposed single-party algorithm, and it has better data utility than the distributed k-anonymity algorithm for classification analysis. In the proposed system we are using the exponential mechanism to choose 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.

Differential Privacy

In this paper, we adopt an exponential mechanism for differential privacy model that provides a provable privacy guarantee. Differential privacy is defined as a staunch model that makes no assumption about an adversary's background knowledge. A exponential private mechanism ensures that the probability of any output is equally similar to the input data sets and, thus, guarantees that all outputs generated out of the private mechanism are insensitive to any Individual's data. In other words, an individual's privacy is not at risk because of the participation in the data set. Data integration methods enable different data providers to flexibly integrate their expertise and deliver highly customizable services. Combining data from different sources could potentially reveal person-specific sensitive information. A secure Distributed k-Anonymity (DkA) framework for integrating two private data tables to a k-anonymous table in which each private table is a vertical partition on the same set of records. Securely integrate private data from multiple parties (data providers). Our algorithm achieves the k-anonymity privacy model in a semi-honest adversary model and employs a game-theoretic approach to thwart malicious participants and to ensure fair and honest participation of multiple data providers in the data integration process.

Architecture Diagram:



Gateway Monitoring:

Gateway monitoring provides security for both the parties in keeping the data private form access to third party users. It also monitors the user request and response for the corresponding request to the particular data set and search is done through the gateway across the validation query. Query results shows the authentication obtained from the both the parties. Each party uses an encryption technique while storing the data on the database and providing to the third party.

Classification Analysis

The proposed algorithm can effectively retain essential information for classification analysis. We evaluate the

scaling impact on the data utility in terms of classification accuracy. Two-Phase Validation algorithm (2PV) is compared with Distributed Differential Exponential Algorithm and with the distributed algorithm for k-anonymity, which refer to as DAKA. The algorithm DAKA integrates and publishes distributed data with k-anonymity privacy guarantee for classification analysis. Finally, we estimate the computation and the communication costs of both the algorithms to provide better security.

Two-Phase Validation (2PV) Algorithm

Random value protocol is used in defining the Two-Phase Validation (2PV) algorithm which operates in two phases: collection and validation. During collection, the Trusted third party first sends a Prepare-to-Validate message to each party server. In response to this message, each party evaluates the proofs for each query of the transaction using the latest policies it has available and sends a reply back to the trusted party containing the truth value along with the data sensitive.

Data Integrity Protection (DIP)

Data Integration Protection scheme is designed under a mobile based adversarial model, which enables a client to feasibly verify the integrity of random subsets of outsourced data against general or malicious corruptions. It takes various parameters for measuring the security-performance of the two parties. DIP scheme can be implemented in a real cloud storage test bed under different parameter choices. We also analyze the security strengths of our DIP scheme via different mathematical models.

Security in Semi-Honest Adversaries

The model that we consider here is that of two-party computation in the presence of static semi-honest adversaries. The protocol specification follows the adversary controls one of the party. A two-party protocol problem is cast by specifying a random process that maps pairs of inputs to pairs of outputs (one for each party).

We refer to such a process as a functionality and denote it f : {0, 1} ∗ ×{0, 1} ∗ → {0, 1} ∗ ×{0, 1} ∗ , where f = (f1, f2).

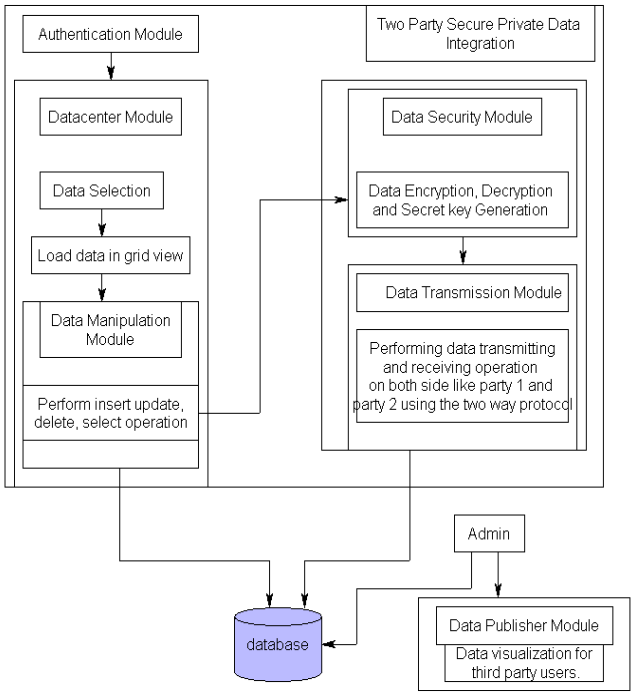
(ie), for every pair of inputs x, y represented as {0, 1}, the output-pair is a random variable (f1(x, y), f2(x, y)) ranging over pairs of strings.

The first party obtains the input x to compute f1(x, y), and the second party obtains the input y to compute f2(x, y). Such a functionality is denoted by (x, y) → (f1(x, y), f2(x, y)). When the functionality f is probabilistic, we denote the notation as f(x, y, r), where r is a uniformly chosen random value used for computing f.

V. IMPLEMENTATION IDEAS

Implementation is the most crucial stage in achieving a successful system and giving the user's confidence that the new system is workable and effective. The objectives of input design are to produce a cost-effective method of input and to achieve a highest possible level of accuracy. The idea is being implemented and is under progress which provides better data utility for classification analysis.

System flow of the entire process is depicted below:



Several activities are carried out to the overall input process that includes,

- Data Authentication & Verification - Collection of data at its source, transfer of data to an input form and conversion of the input data to a computer acceptable medium. After conversion the data is also validated.

- Data Security and Integrity – Data is encrypted and then integrated to be given to the third party for publishing
- Data Control & Transmission – Checking the accuracy and controlling the flow of data to the computer and transferring the data to the computer.
- Data Validation – Checking the input data by program when it enters the computer system.

The main process of implementation includes the following:

Two phase Validation:

Input: Authentication (User Authentication) and Uploading Dataset/Bulk Files
Procedure

1. Validate the authentication corresponding to the dataset
2. Allow Dataset to load into the tables
3. If authentication fails then discard the dataset
4. Datasets are passed on after authentication to create a random value.

Random value protocol:

Input: Dataset/Bulk Files
Procedure

1. Dataset/Bulk Files are partitioned to two subset(public and private)
2. Create random number to store data of different subsets
3. Pass the random value

Data Integration Protection:

Input: Dataset/Bulk Files with Random secret number
Procedure

1. Dataset/Bulk Files are partitioned to two subset(public and private)
2. Partition the dataset into two variable and apply random variable to store the partition data
3. Finally, private attribute are encrypted and stored into database
4. Partitioned datasets are obtained.

Differential privacy:

Input: Authentication and user query
Procedure

1. Verify the authentication type for query to fetch into the database
2. Split the query and execute into the corresponding database based on privacy mechanism
3. Pull the data and validate based on k-anonymity
4. Output value is passed for multiparty computation.

Multiparty computation:

Multi-party computation enables parties to jointly compute a function over their inputs. Joint computation of function $f(x,y)$ without revealing any information about both the values of ‘x’ or ‘y’ exhibits private attributes of data is validated.

We have also proposed a two-phase validation commit (2PVC) to compute the function by integrating the two parties.

Procedure:

Two-Phase Validation Commit - 2PVC

1. Send “Prepare-to-Commit” to all parties
2. Wait for the reply from both the parties
3. If any participant replied No Check or authentication.
4. ABORT the particular party.
5. Identify the authorization identity from both parties.
6. If all party utilize the largest version for each unique data then
7. Integrate the data
8. If any not responded ABORT
9. Otherwise COMMIT
10. end
11. else “integrated data “ for publishing
12. Goto Step 5

VI. CONCLUSION

In this paper, we have presented the concept of Random value protocol used in defining the Two-Phase Validation (2PV) algorithm that operates in two phases with much high security. It provides similar data utility compared to the recently proposed single-party algorithm and better data utility than the distributed k-anonymity algorithm for classification analysis. We expect that the outcome of random value method will be much efficient than the previous methods.

VII. FUTURE WORK

The future research work can to find a solution for using different encryption techniques in order to use it for a multi-party computation. Also, most of the solutions obtained are with respect to a semi-honest adversary model. The future work can be expanded to full security model of computation.

REFERENCES

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

[2] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M.Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 28-34, Dec. 2002.

[3] P. Bunn and R. Ostrovsky, "Secure Two-Party K-Means Cluster-ing," Proc. ACM Conf. Computer and Comm. Security (CCS '07), 2007

[4] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, June 2010.

[5] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[6] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Data Sets," ACM Trans. Database Systems, vol. 33, article 17, 2008.

[7] R.C.W. Wong, J. Li, A.W.C. Fu, and K. Wang, " k-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '06), 2006.

[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining,"J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[9] Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu, "Differentially Private Data Release for Data Mining," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '11), 2011.

[10] J.Vaidya and C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '03), 2003.

[11] N. Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu, "Differentially Private Data Release for Data Mining," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '11), 2011.

[12] R.C.W. Wong, A.W.C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proc. Int'l Conf. Very Large Data Bases, 2007.