A Survey on Privacy Preserving Multiparty Data Mining Techniques

Apoorva Shrivastava[1], Prof. Sandeep Pratap Singh[2], Prof. Sanjay Sharma[3]

[1]M-Tech Research Scholar, [2]Asst Prof, [3]HOD [1, 2, 3] Department of Computer Science and Engineering, OIST, Bhopal

Abstract - The information system is an integrated system that holds financial and personnel records of persons working in various departments. The availability of an increasing amount of user generated data is transformative to our society. However, the large collection of user generated data contains unique patterns and can be used to re-identify individuals. Number of techniques such as privacy preserving data mining, k-anonymity, l-diversity and other techniques are developed for preventing the privacy of data owner and data release operations. In this survey work various techniques are reviewed for privacy preserving data release technique for improving the privacy with minimum cost requirements.

Keywords: Distributed Database, Multi-Party Computation, Aggregation, Privacy Preserving.

1. INTRODUCTION

Data Mining is one of the area gaining lot of practical significance and is progressing at a brisk pace with new methods, methodologies and findings in various applications related to medicine, computer science, bioinformatics and stock market prediction, weather forecast, text, audio and video processing to name a few. Data mining is a process of discovering hidden patterns and information from the existing data.

Increasing network complexity, affording greater access, sharing information and a growing emphasis on the Internet have made information security and privacy a major concern for individuals and organizations. Data mining is a well-known technology for automatically and intelligently extracting knowledge from large amount of data. Such a process, however, can also disclosure sensitive information about individuals compromising the individual's right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting, Privacy preserving data mining (PPDM) is a new era of research in data mining. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy Sensitive data for analysis purposes.

A distributed database for private party information is database in which storage devices are not all attached to a shared processing unit such as the CPU, controlled by a distributed database management system (together sometimes called a distributed database system). It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely-coupled sites that share no physical components. System administrators can distribute collections of data (e.g. in a database) across multiple physical locations [1].

People use telecommunication-based applications daily and the system collects a large amount of information related to their activities. Such telecom networks create opportunities for joint cooperative tasks based on computations with inputs supplied by separate users. Moreover, such computations could be performed even among mutually untrusting parties. Since many user inputs are private information reflecting users' daily activities (for example travel routes, buying habits, and so on), secure multi-party computations (SMC) become a relevant and practical approach for dealing with such applications. Many applications can utilize available private data to improve the quality and security of everyday life [2].

Secure Multi-party Computations and Privacy

Internet and distributed computer architecture provides innumerable possibilities for collaborative/joint computations. SMC is a mechanism for privacy preserving data mining which is meant for joint computations in networked environment. It can be defined as, to provide computations among several diverse organizations in a safe or secure manner. With SMC, several parties can jointly perform some global computation on their private data without any loss of data security/privacy. It provides base for end-to-end secure multiparty protocol development.

Many SMC-based solutions can be used to ensure privacy preservation and protection. Informally, they can be described as a computational process where two or more parties compute a function based on private inputs. Privacy in this context means that none of the parties wants to disclose its own input to any other party [2, 3].

A common architecture for database access is client-server, where the server manages the data and answers clients' queries according to its access policy. In such architecture, there may be two very distinct privacy considerations. The first has to do with client's privacy and is highly motivated in cases where the server's knowledge of client's queries may be harmful, for instance, in patent litigation or market research. In such cases, without an expectation of privacy, clients may be discouraged from querying the database in the first place. Such concerns can be answered using various cryptographic solutions such as oblivious transfer, single-server private-information retrieval, and more generally, secure function evaluation (SFE), which may be used to restore privacy for the clients [4].

Huge 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 [5].

Thus the proposed work is a privacy management and data security scheme that employed over the centralized data aggregation model and able to produces the data release with adding some amount of noise for securing the sensitivity of actual data among different parties. The databases consume the centralized approaches to reduce the complexity of data. But data in same place can generate the issues during the data access thus the proposed work is data privacy preserving, data analysis and release technique is simulated using vertically partitioned data.

2. BACKGROUND

Differentially Privacy

Differential privacy will take the view that it was not, with the rationale that the impact on the smoker is the same independent of whether or not he was in the study. It is the conclusions reached in the study that affect the smoker, not his presence or absence in the data set. Differential privacy ensures that the same conclusions, for example, smoking causes cancer, will be reached, independent of whether any individual opts into or opts out of the data set. Specifically, it ensures that any sequence of outputs (responses to queries) is "essentially" equally likely to occur, independent of the presence or absence of any individual. Here, the probabilities are taken over random choices made by the privacy mechanism (something controlled by the data curator), and the term "essentially" is captured by a parameter. A smaller will yield better privacy (and less accurate responses). Differential privacy is a definition, not an algorithm. For a given computational task T and a given value of " there will be many differentially private algorithms for achieving T in an "-differentially private manner. Some will have better accuracy than others. When " is small, finding a highly accurate "-differentially private algorithm for T can be difficult, much as finding a numerically stable algorithm for a specific computational task can require effort [6]. Differential privacy (DP) is a well-studied notion of privacy that is generally achieved by randomizing outputs to preserve the privacy of the input records.

Differential privacy mechanisms work by randomizing the output to preserve the privacy of the input records. Thus the main question in differential privacy is *quantitative* rather than *qualitative*: How much accuracy must be lost in order to preserve input privacy.

Privacy Preserving Scientific Computations

Many industries, including the telecommunication industry, have to solve problems related to planning, routing, scheduling or optimization. These problems are often modeled as systems of linear equations or linear least squares problems. However,

in the scenario in which two or more untrusted parties want to solve the problems without revealing private data, traditional well-studied approaches are not applicable [2].

The subsequent real-life state of affairs additional illustrates the requirement for synchronal knowledge sharing and privacy preservation of person-specific sensitive knowledge.

Data can be horizontally-partitioned among different parties over the same set of attributes. These distributed data can be integrated for making better decisions and providing high-quality services. However, data integration should be conducted in a way that no more information than necessary should be revealed between the participating entities. At the same time, novel knowledge that results from the integration process should not be misused by adversaries to

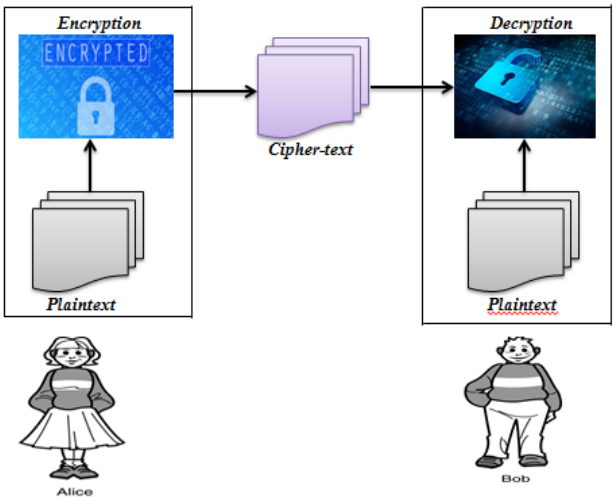

Figure 1 Secure Two Party Communications

reveal sensitive information that has not been available before the data integration. The challenge in data privacy is to share data while protecting personally identifiable information. Differential privacy is a strong privacy definition. It is a sub protocol of main algorithm. It uses the exponential mechanism in a distributed setting. Two party data publishing algorithm for vertically partitioned data that generate an integrated data table satisfying differential privacy. Distributed data can be integrated to enable better data analysis for making better decisions and providing high-quality services. The first two-party differentially private data release algorithm for vertically partitioned data. It respect to a utility function while preserving differential privacy. A secure mechanism is required to compute the same output while ensuring that no extra information is leaked to any party [7].

3. LITERATURE SURVEY

The given section provides the study of recently developed approaches which are used for improving the security and sensitivity during the data aggregation and release.

Current advances in information, communications, data mining, and security technologies have gave rise to a new era of research, known as Privacy Preserving Data Mining

(PPDM). Privacy preserving data mining has become increasingly popular because it allows sharing of privacy Sensitive data for analysis purposes. Several data mining algorithms, incorporating privacy preserving mechanisms, have been developed that allow one to extract relevant knowledge from large amount of data, while hide sensitive data or information from disclosure or inference. Finally, we share directions for future research. In this paperTamanna *Kachwala* Sweta Parmar* [14] provides a review of the state-of-the-art methods for privacy and analyze the representative technique for privacy preserving data mining

Data mining aims to take out useful information from multiple sources, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. Privacy preserving data mining (PPDM) is a new research direction in data mining and statistical databases , where data mining algorithms are analyzed for the side effects they acquire in data privacy. The main consideration of the privacy preserving data mining is twofold. First, sensitive raw data like identifiers, name, addresses and the like should be modified out from the original database, in order for the recipient of the data not to be able to cooperation another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well cooperation data privacy. The main purpose of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge stay private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies. Agarwal and Srikant and Lindell and Pinkas introduced the first Privacy -preserving data mining algorithms which allow parties to collaborate in the extraction of knowledge, without any party having to reveal individual items or data. The goal of this paper is to give a review of the different dimension and classification of privacy preservation techniques used in privacy preserving data mining. Also aim is to give different data mining algorithms used in PPDM and related research in this field.

Recent advances in information, communications, data mining, and security technologies have gave rise to a new era of research, known as Privacy Preserving Data Mining (PPDM). Several data mining algorithms, incorporating privacy preserving mechanisms, have been developed that allow one to extract relevant knowledge from large amount of data, while hide sensitive data or information from disclosure or inference. PPDM is a new attempt; thus, several research questions have often being asked. For instance: (1) How to measure the performance of these algorithms? (2) How effective of these algorithms in terms of privacy preserving? (3) Will they impact the accuracy of data mining results? and (4) Which one can better protect sensitive information? To help answer these questions, we conduct an extensive review on literature. We present a classification scheme, adopted from early studies, to guide the review process.

PPDM has recently emerged as a new field of study. As a new comer, PPDM may offer a wide application prospect but at the same time it also brings us many issues /

problems to be answered. In this paper. *Srinivasa Rao & B. Srinivasa Rao* [13] conduct a comprehensive survey on 29 prior studies to find out the current status of PPDM development. We propose a generic PPDM framework and a simplified taxonomy to help understand the problem and explore possible research issues. We also examine the strengths and weaknesses of different privacy preserving techniques and summarize general principles from early research to guide the selection of PPDM algorithms. As part of future work, we plan to apply the proposed evaluation framework to formally test a complete spectrum of PPDM algorithms.

The availability of an increasing amount of user generated data is transformative to our society. We enjoy the benefits of analyzing big data for public interest, such as disease outbreak detection and traffic control, as well as for commercial interests, such as smart grid and product recommendation. However, the large collection of user generated data contains unique patterns and can be used to re-identify individuals, which has been exemplified by the AOL search log release incident. In this paper, Liyue Fan et al [8] propose a practical framework for data analytics, while providing differential privacy guarantees to individual data contributors. Given framework generates differentially private aggregates which can be used to perform data mining and recommendation tasks. To alleviate the high perturbation errors introduced by the differential privacy mechanism, authors present two methods with different sampling techniques to draw a subset of individual data for analysis. Empirical studies with real world data sets show that solutions enable accurate data analytics on a small fraction of the input data, reducing user privacy risk and data storage requirement without compromising the analysis results.

Statistics from security firms, research institutions and government organizations show that the number of data-leak instances has grown rapidly in recent years. Among various data-leak cases, human mistakes are one of the main causes of data loss. There exist solutions detecting inadvertent sensitive data leaks caused by human mistakes and to provide alerts for organizations. A common approach is to screen content in storage and transmission for exposed sensitive information. Such an approach usually requires the detection operation to be conducted in secrecy. However, this secrecy requirement is challenging to satisfy in practice, as detection servers may be compromised or outsourced. In this paper, XiaokuiShu et al [9] present privacy-preserving data-leak detection (DLD) solution to solve the issue where a special set of sensitive data digests is used in detection. The advantage of our method is that it enables the data owner to safely delegate the detection operation to a semi-honest provider without revealing the sensitive data to the provider. Authors describe how Internet service providers can offer their customers DLD as an add-on service with strong privacy guarantees. The evaluation results show that our method can support accurate detection with very small number of false alarms under various data-leak scenarios.

In this paper, AlevtinaDubovitskaya et al [10] address the problem of building an anonym zed medical database from multiple sources. The given solution defines how to achieve data integration in a heterogeneous network of many clinical institutions, while preserving data utility and patients' privacy. The contribution of the paper is twofold: Firstly, a secure and scalable cloud e-Health architecture to store and exchange patients' data for the treatment. Secondly, an algorithm for efficient aggregation of the health data for the research purposes from multiple sources independently.

Protecting the privacy of individuals in graph structured data while making accurate versions of the data available is one of the most challenging problems in data privacy. Most efforts to date to perform this data release end up mired in complexity, overwhelm the signal with noise, and are not effective for use in practice. In this paper, Jun Zhang et al [11] introduce a new method which guarantees differential privacy. It specifies a probability distribution over possible outputs that are carefully defined to maximize the utility for the given input, while still providing the required privacy level. The distribution is designed to form a 'ladder', so that each output achieves the highest 'rung' (maximum probability) compared to less preferable outputs. Authors show how ladder framework can be applied to problems of counting the number of occurrences of sub-graphs, a vital objective in graph analysis, and give algorithms whose cost is comparable to that of computing the count exactly. The experimental study confirms that given method outperforms existing methods for counting triangles and stars in terms of accuracy, and provides solutions for some problems for which no effective method was previously known. The results of algorithms can be used to estimate the parameters of suitable graph models, allowing synthetic graphs to be sampled.

Computer forensics and privacy protection fields are two conflicting directions in computer security. In the other words, computer forensics tools try to discover and extract digital evidences related to a specific crime, while privacy protection techniques aim at protecting the data owner's privacy. As a result, finding a balance between these two fields is a serious challenge. Existing privacy preserving computer forensics solutions consider all data owner's data as private and, as a result, they collect and encrypt the entire data. This increases the investigation cost in terms of time and resources. So, there is a need for having privacy levels for computer forensics so that only relevant data are

collected and then only private relevant data are encrypted. WaleedHalboob et al [12] propos privacy levels for computer forensics. It starts with classifying forensic data, and analyzing all data access possibilities in computer forensics. Then, it defines several privacy levels based on the found access possibilities. The defined privacy levels lead to more efficient privacy-preserving computer forensics solution.

4. CONCLUSION

Differential privacy is a strong privacy mechanism. However, Review and study of research results have shown that if the records are not independent or an adversary has access to aggregate level background knowledge about the data, then privacy attack is possible. In this brief review of the existing research, we considered the main idea behind secure multi-party computations and how they can be used to develop distributed applications that preserve the privacy of participating parties. The algorithms in this paper guarantees privacy, preserve data utility for data mining, and are scalable for handling large data sets.

REFERENCES

[1] P.Lalitha, D.BullaRaoandP.NageswaraRao, "Novel Secure Multiparty Protocol in DistributedDatabases", *International Journal of Innovative Research in Computer and Communication Engineering*,Vol.2, September 2014.

[2] Vladimir A. Oleshchuk and Vladimir A. Dorozhny, "Secure Multi-party Computations and Privacy Preservation:Results and Open Problems", Telektronikk 2.2007

[3] Dr. Durgesh Kumar Mishra and PurnimaTrivedi, "A Glance at Secure Multiparty Computation forPrivacy Preserving Data Mining", *International Journal on Computer Science and Engineering*, Volume, PP. 171-175, 2009.

[4] Andrew McGregor and IlyaMironov, "The Limits of Two-Party Differential Privacy", *Foundations of Computer Science (FOCS), 51st Annual IEEE Symposium*, PP. 81 – 90, 2010, Las Vegas, NV.

[5] Ms. Gauri V. Sonawane and Prof. NareshThoutam, "Design And Analysis Of Two Party Protocol to RetrievePrivate Data By Using Vertical Partitioned Data-A Review",*International Journal of Electronics, Communication & Soft Computing Science and Engineering*, Volume 3, PP. 21 – 24.

[6] K.Samhita and M.Sannihitha, "ϵ-Differential Privacy Model ForVertically Partitioned Data ToSecure The Private Data Release", *International Journal of Technical Research and Applications*, Volume 3, (May-June 2015), PP. 88-93.

[7] TAMILMANI.M and BALAMURUGAN.D, "Multiparty Private Data Publishing Using Exponential Trust Based Mechanism in Malicious Adversary Model", *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, Volume 14, APRIL 2015.

[8] Liyue Fan, Hongxia Jin, "A Practical Framework for Privacy-Preserving Data Analytics", May 18–22, 2015, Florence, Italy, ACM.

[9] XiaokuiShu, Danfeng Yao, and Elisa Bertino, "Privacy-Preserving Detection of Sensitive Data Exposure", *IEEE Transactions on Information Forensics and Security*, VOL. 10, NO. 5, MAY 2015.

[10] AlevtinaDubovitskaya, VisaraUrovi, MatteoVasirani, Karl Aberer, and Michael I. Schumacher, "A Cloud-Based eHealth Architecture for Privacy Preserving Data Integration", *IFIP International Federation for Information Processing 2015*, IFIP AICT 455, pp. 585–598, 2015.

[11] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, DiveshSrivastava, Xiaokui Xiao, "Private Release of Graph Statistics using Ladder Functions", *SIGMOD'15*, May 31–June 4, 2015, Melbourne, Victoria, 2015 ACM.

[12] WaleedHalboob, RamlanMahmod, NurIzuraUdzir, Mohd. Taufik Abdullah, "Privacy Levels for Computer Forensics: Toward a More Efficient Privacy-preserving Investigation", *International Workshop on Cyber Security and Digital Investigation (CSDI 2015)*Elsevier

[13] *K. Srinivasa Rao & B. Srinivasa Rao ," An Insight in to Privacy Preserving Data Mining Methods*" The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 3, July-August 2013

[14] *Tamanna Kachwala* Sweta Parmar***, "***An Approach for Preserving Privacy in Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering.*