Vol.4,No.1, July, 2015
44 - 49,ISSN:0975-5446] [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

Privacy, Collaboration and Security – Imperative Existence in Data Mining

ALPA SHAH
MCA Department
Sarvajanik College of Engineering and Technology, Surat, India
alpa.shah@scet.ac.in

RAVI GULATI
Department of Computer Science,
Veer Narmad South Gujarat University,
Surat, India
rmgulati@gmail.com

Abstract

Privacy and security can be impediment for data mining carried out in collaboration. A clear demarcation between security and privacy of information provided by the contributing parties must be determined. This paper addresses the concern of identifying the importance of security and privacy in data mining. Important aspects of security and privacy with collaborative data mining are also conferred.

Keywords: data mining, collaboration, cryptographic primitives, privacy preserving data mining

1. Introduction

An unprecedented growth for data collection and its analysis with the use of technology has been witnessed in recent years. Both organizations and individuals generate a large pool of data with routine activities. The data is often stored as a repository at some central location or mined for pattern recognition at some distributed site. Data Mining is used intensively to extract useful information from this vivacious data generated. Data mining extracts knowledge supporting a large domain of marketing, medical diagnosis, weather forecasting and national security. Emerging research has contributed to formulate algorithms to extract the meaningful data from the enormous pool of data. Digital collection system for data has increased a concern for privacy amongst the individuals.

Privacy is a socio-cultural concept. With latest emerging trends of Web and ubiquity of computers privacy is now a digital issue. In a data mining system, disclosed private information (from one entity to another) should be the minimum necessary for data mining [10][4]. Basically,

data disclosure among peer entities governs three fundamentals for privacy. *Data collection based protocols* protects privacy during data transmission from the data providers to the data warehouse server. *Inference control based protocols* manages privacy protection between the data warehouse server and data mining servers. *Information sharing based protocols* controls information shared among the data mining servers in different systems. A common goal of these protocols is to leverage minimized information loss and maximized privacy of participating entities. A crucial challenge to mine definite data without violating the privacy of individual is still a wooly covered issue.

2. Desire for Collaboration

With increase in competitive market, parties have started sharing their rich datasets of information about their consumers and their buying habits. Such parties use data mining to identify patterns/trends from data collected from local distributed sites. Such collaborative computation is desirable by competitive peers to enhance the mutual benefit of each of it. The parties desire the privacy of their customers, but understand the gain of mutual collaboration. The contributing parties use data mining technologies to identify patterns/trends from the shared data. In last few years there is an increased demand for collaborative mining over distributed data to conjecture rules for the benefit of competitive parties. This mutual contribution by the competitive parties has opened a new area of research termed "Privacy Preserving Collaborative Data Mining [PPCDM]" [7] [8]. A lot of privacy issues have been raised in context of data mining by collaboration.

The importance of PPCDM can be compelling under several scenarios. Private hospitals might want to find the prevailing trends of disease at diverse locations. Effective mining amongst peer hospitals is desirable, without violating the privacy concern of its personal patient's records. Supermarkets might want to collaborate to infer the buying trends of their customers. Each participating supermarket doesn't want to disclose much about their customer's information, but still understand magnitude of such joint data mining activity. Anti terror organizations, surveillance systems might want to collaborate such activities for mutual benefits. Various pharmaceutical companies might want to collaborate to find meaningful patterns on compounds used for medicinal purposes.

The need for privacy is thus sometimes due to public welfare like medical databases or surveillance systems or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it is scientific or economic and market oriented. The main challenge of PPCDM is to conduct effective mining without possible loss of privacy of data from contributing parties.

3. Security Aspect of Data mining

Cryptographic techniques [1] have been developed to protect the data at a central site or during the transmission. The model assumes semi-honest but curious participants [6]. The parties don't deviate from the protocols but can provide false inputs so that they can gain advantage. The crux being, two parties motive to conduct data mining based on private inputs, but neither party willing to disclose its input to other. The parties can change their inputs in order to gain more

information about the inputs of other parties. A party can also deviate by sending empty input set. A computation can be done which will now generate output resulting from only the other party. A malicious adversary may arbitrarily deviate from the protocol specification itself.

Homomorphic encryptions are based on public key encryption and decryption techniques. In Homomorphic encryption a sender can manipulate cipher texts that encrypt data under some public key p_k to construct a cipher text that encrypts 'any desired function' of that data under p_k [11]. Homomorphic encryption schemes are a special class of public key encryption schemes. Homomorphic encryption has proved effective in defending against semi honest adversaries [3]. In this method, each data mining server encrypts its local data mining model and exchanges the encrypted model with other data mining servers. Some encryption scheme properties, such as the Rivest- Shamir-Adleman (RSA) cryptosystem's commutative encryption property, make it possible to design algorithms for data mining servers to perform certain data mining tasks.

Oblivious transfer is a mechanism in which sender divides the information in potentially n different ways and sends it to the receiver, remaining oblivious with what piece of information he is sending. The first form of oblivious transfer that is 1-out-2 is widely used for secure computation without the use of any other cryptographic primitives. Oblivious transfer is the most computationally intensive operation of secure protocols being repeated many times. Each invocation of oblivious transfer typically re- requires a constant number of invocations of trapdoor permutations (i.e. public-key operations, or exponentiations).

The novel idea of secret sharing was developed by Shamir [5] and Blakley[12] in 1979. The idea is that one party has a secret which it distributes among n other parties in a way that none of the n parties alone can recover the secret. The design is such that the information of at least *i* out of *x* parties is needed to recover the secret. Here *i* is defined as the threshold for which parties are required to get the secret. Any attempt for less than i parties will result into failure of recovering the secret.

Sharing a secret key is based on the potential of the key itself. Higher the power of key more it is susceptible for security. But most of the systems require the secret to be broken down into a minimal of say n inputs. This may sometimes be unsuitable for applications where there are fewer hosts say 2 or 3. Another drawback is that when two or more of the parties deliberately behave dishonestly to gather some more information. A variant, called verifiable secret sharing adds extra information to each secret but which increases communication overhead cost of the protocol.

4. Privacy Aspect of Data Mining

Preserving the privacy of data of the contributing party is studied extensively by the data mining community. Privacy preserving data mining algorithms can be broadly classified in three categories:

1. *Data distortion based privacy*: These algorithms aim at distorting the original private data, when released, do not divulge any individually identifiable information. Data perturbation

privacy preservation techniques [9] aim at modifying the private data values by adding additive or multiplicative noise drawn from a probability distribution to the data values.

2. Cryptography based privacy: Cryptographic protocols are called private when their execution does not reveal any additional information about the involved parties' data, other than what is computed as a result of the protocol execution.

3. Output perturbation based privacy: Output perturbation techniques [12] discuss privacy with respect to the information released as a result of querying a statistical database by some external entity. Output perturbation based techniques, do not identify specific data attributes to be more privacy sensitive than others. Privacy is achieved by defining algorithmic mechanisms called sanitizers that work by perturbing the output of a query function on the database.

5. Considerations For Security and Privacy for Data Mining

Data mining confronts challenges for security of data at the central repository or distributed collaborating sites. The level of security requirements can be enforced by the individual parties or by

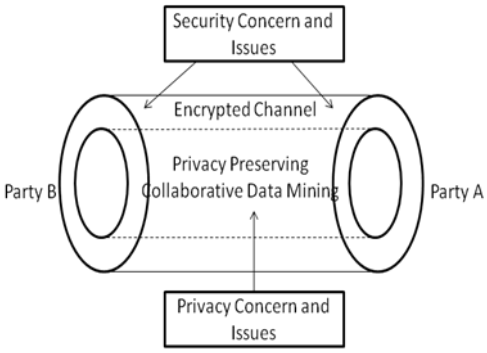

Figure 1: Security and Privacy Aspects in PPCDM

Legislative laws. Cryptographic tools preserve the output of computations. They can be used to prevent privacy leaks in the process of computations. They can also be used while data is transferred from local sites to some central site for mining. Another important aspect during mining is privacy of data at local sites. Privacy of data must be confronted by law enforcement at local collecting sites. Data can be secured with encryption or envelope techniques. At the central site various techniques for maintaining privacy like perturbation, secure sum computation , anonymization should be used. For PPCDM, each collaborating parties must ensure the privacy level. Also, each of the collection center as well as the data repository center must be secured. Privacy protection can be conferred to provide two important criteria. Firstly, specifying different levels of privacy requirements which need to be satisfied by each contributing parties. Also, for attribute sensitive data, concern for uniform attributes to be considered sensitive. Say, a party may consider birth date to be private information while other party might not consider it private attribute and disclose it publicly.

The performance of PPCDM protocols are measured based on three parameters: effectiveness of results, efficiency of protocols and privacy degree achieved by applying the protocol. PPCDM protocols devised with different techniques have different performance indicators based on the parameters described. Some algorithms are more efficient and secure, the security aspect, but the joint data mining results are greatly affected, privacy aspect. Others produce accurate outcomes and absolute security, but the computational or communicational complexity is too high to be accepted. In PPCDM, decision-making strategy is required to analyze the best fit solution for the proposed business problems. The trade off exists between the privacy of the individual information and the correctness of the data mining results. That is, privacy is achieved at the cost of accurate outcome.

A proposed scheme to formulate a privacy preserving model confronting the security aspects is discussed hereby. The scheme formulates a privacy preserving model confronting the security aspects of PPCDM. Figure 1 describes the core of the scenario. The idea is to demarcate the requirements in privacy and security of PPCDM. The scheme is summarized for two-party case. It can be extended for n-parties desiring for mutual collaboration. Parties A and B want to collaborate for mutual benefits. The model assumes that the parties share their data with each other without a trusted central repository. Within the secured channel, when data is exchanged for mutual collaboration, encryption based cryptographic techniques are fundamental. This is required to counteract leakage of data to a malicious user. Active attacks can be forfeited with proper cryptographic techniques. Also each collaborating party can exchange their data using Oblivious Transfer Protocol. It is important to address the security aspects and concern of data during this stage of PPCDM. After the data is available at distributed site, mining of data to extract meaningful information is desirable. Privacy level requirement should be analyzed. Various privacy preserving algorithms should be used to sustain the privacy constraints of the parties involved.

6. Conclusions

The field of collaborative data mining has seen considerable research in the last decade. Participating entities range from individual users to governmental and transnational organizations. This makes data privacy a primary concern in the deployment of such applications in the real world. Cryptographic techniques can be used for dealing with privacy issues in distributed computing environments. The privacy concerns of the different participating entities vary, as does their ability to protect their private data due to varying availability of resources. Monolithic privacy models entrust a leveled privacy for all the participating entities. A desirable model for PPCDM must provide privacy stratum based on the requirement of the participations. Perturbation and secured sum computations based methods can be used for preserving the privacy of the participants.

Acknowledgement

We acknowledge UGC for Special Assistance Program (SAP) for Natural Language Processing and Data Mining (file number is F.3-48/2011) under which this work has been done.

References

1. Benny Pinkas," *Cryptographic techniques for privacy preserving data mining" in* SIGKDD Explorations. Volume 4, Issue 2 - page 12, 2002
2. Y. Lindell and B. Pinkas "*Privacy preserving data mining"* in advances in Cryptology (CRYPTO'00), volume 1880 of Lecture Notes in Computer Science, pages 36-53, Springer-Verlag, 2000.
3. C Gentry, S Halevi, "*Implementing Gentry's Fully-Homomorphic Encryption Scheme* ", Advances in Cryptology--EUROCRYPT 2011, Springer.
4. Chris Clifton and Donald Marks "*Security and privacy implications of data mining"*, in Proceedings of the ACM SIGMOD Workshop o n Research Issues on Data Mining and Knowledge Discovery.
5. A. Shamir "*How to share a secret"* in Communications of the ACM, 22(11):612–613, November 1979.
6. Keke Chen and Ling Liu "*Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation*" , Parallel ad Distributed Systems, IEEE Transactions, Vol 20, No. 12, December 2009
7. Justin Zhan "*Privacy Preserving Collaborative Data Mining*" , IEEE Computational Intelligence Magazine, May 2008 Keke Chen;
8. Ling Lu Vassilios S, Elisa, Igor Nai, Loreana P, Yucel, Yannis T, "*State-of-art In Privacy Preserving Data Mining*" SIGMOD Record, Vol 33, No1 March 2004
9. W. Du and Z. Zhan, "*Using randomized response techniques for privacy-preserving data mining*" In Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, Aug. 24–27, 2003.
10. Nan Zhang and Wei Zhao "*Privacy-Preserving Data Mining Systems*" in magazine published by IEEE Computer Society 2007.
11. G. R. Blakley, "*Safeguarding cryptographic keys*" in Proceedings of National Computer Conference, pages 313–317, June 1979.
12. Yehuda Lindell and Benny Pinkas "*Secure Multiparty Computation for Privacy-Preserving Data Mining*" in The Journal of Privacy and Confidentiality (2009), Number 1, pp. 59–98