A Report of the Privacy in Data Mining: Speakers Survey

Ramana Bonathu, Devaki, K. Ramalinga Reddy, Dhasaratham Meghavath, G. Vijaya

Abstract: we study the data mining understand the need for analyses of large, complex, information rich data sets. Privacy is mining of distributed data has numerous applications. The privacy preserving data mining has been developed, understanding the role of privacy in data mining is difficult. Many algorithms and approaches that have been developed theoretically, but practically it is difficult. In this paper specify the overview the privacy preserving in data mining. This paper mainly focuses on the literature survey of the data mining privacy. Here many speakers tell to their own ideas, these ideas to develop new standards and privacy algorithms in data mining.

Keywords: Data mining, Data privacy ,privacy preserving

I. INTRODUCTION

This paper is concerned with the study and analysis of preserving privacy in collaborative data mining in order to improve the efficiency and effectiveness of privacy preservation among the collaborative parties. Data mining is a process which uses different data analysis tools that discover patterns and relationships in data that can be used to make predictions (Sybil and Liebeman, 2001). Most existing data mining algorithms are carried out under the assumption that all the data could be available at single central site. While two or more parties, who don't have enough confidence in each other or even adversaryindividuals or organizations, have a common desire to extract knowledge from all of their private data, the privacy problems come up. As the data mining in the public and private sectors is increasingly used, privacy is becoming an important issue. When common users are involved in data mining, all users need to send their data to trusted common centre to conduct the mining; however, in situations with privacy concerns, it is very difficult for a user to trust the other users and in such a situation, the process is called Privacy Preserving Collaborative Data Mining(PPDM) (Yasien, 2007) and the gap between the data mining and data confidentiality can be filled by the privacy preserving data mining. The trusted partners share information between themselves, by maintaining their privacy, with the secure cooperative computations. The shared information sent by the participant to a remote database also contains privacy data inferences which should not be disclosed to the receiving participant. In order to maintain privacy of individual's, when sharing data in public domain, needs to be secured.

Manuscript received March, 2014.

Dr.Ramana Bonathu, Professor, HOD,Department of CSE,Vidya Vikas Institute of technology,Hyd, India.

Devaki, Lecturer,Dept of CSE, Singareni Colleries Women's Degree &PG college., India.

K.Ramalinga Reddy, Assistant Professor,Dept of CSE, A.M.Reddy college of Engineering, India.

Dhasaratham Meghavath, Lecturer inCSE Dept, Bule Hora University, India.

G.Vijaya, Lecturer in CSE,Dept of CSE,Singareni Colleries Women's Degree & PG college, India.

Few constraints and computations are required to maintain in the public domain during the information sharing happen between the participants.

II. LIMITATIONS ON RESULTS

How can we constrain the results of data mining? There has been work in this area, addressing specific problems such as hiding specific association rules [3, 20] or limiting the confidence in any data mining[5]. While these provide some specific techniques, the means available to constrain results can are quite limited. What is needed is a general way to specify what is and is not allowed. One possible approach is constraint-based data mining[4]. This line of research is concerned with improving the efficiency of algorithms and understandability of results through providing up-front constraints on what result would be of interest. Would the languages used to describe these constraints also serve to define what results are acceptable from a privacy standpoint? While the current methods do not enforce that nothing outside the constraints can be learned, they could provide a starting point for further research. The rest of this paper provides some specific suggestions
for methods to specify privacy constraints in ways that still allow data mining. We start with a discussion of individual privacy, and methods to protect it. We then discuss methods for corporate privacy, or constraining what is learned from a collection. We conclude with several orthogonal metrics for defining and measuring privacy.

III. OVERVIEW OF SECURITY AND PRIVACY IN DATA MINING

Here we can discuss the overview of security and privacy in detaining. here we are taking the various speakers approaches, these approaches are concerned that what are the overview particular topic and specify the their recommendations.

3.1. Chris Clifton - Is Privacy Still an Issue for Data Mining?

overview: The speaker first gave a brief review of the history of PPDM. He pointed out that although PPD research has been active in academia, there are still no practical applications in industry. One reason for this is the lack of understanding about what the privacy related problems are and how they relate to data mining. Understanding the problem is critical for marketing the technology. The real problem, emphasized by the speaker, is the misuse of data. For example, card systems usually save customers' data for analysis; however, without data mining, storing those data long term is not necessary. Therefore, data mining is a cause of data misuse and PPDM can help address this problem.

As a result, the speaker suggested marketing PPDM as a means of protection against misuse. The speaker also discussed the possibility of marketing PPDM as a collaboration technology, e.g., secure supply chain management. Finally, the speaker identified some key issues for the next generation of PPDM (to be described in the next subsection).

Recommendations:

- Develop a formal and practical definition of privacy. It is not only associated with individually identifiable data.
- Develop PPDM techniques that support profitable usage, e.g., controlling disclosure risk/cost, optimizing supply chain without losing competitive advantage, etc.
- Understand the benefits of data mining. How do we measure the confidence in data mining results?
- How do we limit an adversary's learning ability? Can privacy be incentive based? For example, are people willing to give better data if privacy is protected?

3.2. Alessandro Acquisti and Ralph Gross - Privacy Risks for Mining Online Social Networks

overview: The research presented focuses on privacy risks associated with information sharing in online social networks. Online social networks including Facebook, Friendster, and MySpace have grown exponentially in recent years. However, because participants reveal vast amounts of personal and sometimes sensitive information, these computer-mediate social interactions raise a number of privacy concerns.

In an effort to quantify the privacy risk associated with these networks, the authors combined online social network data and other publicly available data sets in order to estimate whether it is possible to re-identify PII (personally identifying information) from simple PI (personal information). This research supports the claim that large amounts of private information are available publicly.

Recommendations:

- Identify ways to quantify the degrees of privacy associated with publicly available data and information shared in onlinesocial networks.
- Develop efficient mitigation strategies that can enhance privacy while preserving valuable online interactions.

3.3. Jaideep Srivastava - Extraction and Analysis of Cognitive Networks from Electronic Communication

overview: Social network analysis focuses on understanding social relationships and interactions within a group of individuals. Cognitive analysis of social networks focuses on understanding what an individual's perception is about other individuals in the network. The speaker began by modeling cognitive social networks and presenting quantitative measures for perception and belief. He then illustrated the usefulness of these ideas using the Enron email communication network and also attempts to identify concealed relationships. The speaker then discussed the problem of modeling and analyzing group dynamics in a social network. A new domain for analyzing group dynamics is massively multi-player online games that include tens of thousands of players who work together in groups to accomplish tasks within the game. While this data set, extracted from web logs, is well-suited for understanding the dynamics of group behavior, data collection, appropriate mining algorithms, and scalability are large issues. Further, the theoretical framework for group behavior is still in its infancy, particularly for adhoc groups.

Recommendations:

- Support interdisciplinary research that will advance computer science as well as other disciplines.
- Develop new, scalable approaches for data access and data cleaning.
- Encourage use of large, real world data sets to validate new data mining algorithms.
- While security is a necessity, a balance between PPDM and data analysis is necessary. Future research need to consider information flow during data analysis.

3.4. Lisa Singh - Exploring Graph Mining Approaches for Dynamic Heterogeneous Networks

overview: Much graph mining research to date focuses on simple network models containing a single node type anda single edge type. In this talk, the speaker discussed the need to develop hidden community identification, spread of influence, and group formation mining algorithms for graphs involving many different node and edge types. Because these graphs are large, graph approximations are necessary to adequately tackle different graph mining problems. The speaker described different approximations and abstractions of complex networks for prediction, visualization, and privacy in the context of observational scientific data. Questions under investigation include: when should we use attributes vs. link structure when building predictive models, how can we use visualization to enhance the quality of mining results, and can we use the same abstraction for different mining applications?. In the context of privacy, the speaker discussed the need to formally define what constitutes a privacy breach within a graph. To date, researchers have proposed conflicting definitions. She then discussed the need to understand network topology in order to effectively determine when certain abstractions of the graph are more private than others.

Recommendations:

- Promote developing graph mining algorithms for complex, dynamic networks with multiple node and edge types.
- Define privacy breaches in the graphs. What constitutes a breach?
- Develop metrics for understanding the topology of graphs? This structure can then be used to measure the level of anonymity in the network.
- Consider the privacy questions in the context of complex, not simple networks.

IV. PERFECT PRIVACY

One problem with the above is the tradeoff between privacy and accuracy of the data mining results. Can we do better? Using the concept of Secure Multiparty Computation, the answer is clearly yes – in the "web survey" example,the respondents can engage in a secure multi party computation to obtain the results, and reveal no information that is not contained in the results. However getting thousands of respondents to participate synchronously in a complex protocol is impractical.

While useful in the corporatemodel, it is not appropriate for the web model. Here we present a solution based on a moderately trusted third party the party is not trusted with exact data, but trusted only not to collude with the "data receiver". There are various means of achieving privacy, both technical and nontechnical. Part of the problem is the need to create a solution which is feasible in terms of efficiency, security, and without limitations in usability. Technical solutions

can be formulated without restrictions in usability, by making suitable assumptions. By a judicious use of nontechnical mechanisms, we can realize these assumptions in real life.

Perfect privacy in the SMC sense implies that there is absolutely no release of any meaningful information to any third party. Current e-commerce transactions have a trusted(central) third party with access to all the information. The "trust" is governed by legal contracts enjoining the improper release of information. In some cases, the third party is dispensed with and contracts exist between the interested parties themselves. This is obviously insecure from the technical perspective. Though it has been proven that aSMC solution does exist for any functionality, the computation and/or communication required may be high. Other factors, such as the need for continual online availability of the parties, create further restrictions and problems in real world settings such as a web-based survey. However, if we jettison the idea of using only the interested

parties, we can obtain a middle ground solution that does not require a fully trusted third party. We can instead use a fixed number of un trusted, non colluding parties/sites to do the computation. Assume the existence of k untrusted, noncolluding sites. Untrusted implies that none of these sites should be able to gain any useful information from any of the inputs of the local sites. Noncolluding implies that none of these sites should collude with any other sites to obtain information beyond the protocol.

Then, all of the local parties can split their local inputs intok random shares which are then split across the k un trusted sites. Each of these random shares are meaningless information by themselves. However, if any of the parties combined their data, they would gain some meaningful information from the combined data. For this reason, we require that the sites be noncolluding. We believe this assumption is not unrealistic. Each site combines the shares of the datait has received using a secure protocol to get the required data mining result. The following is a brief description of this approach. Every party is assumed to have a single bit of information xi, identified by some key i. Each party locally generates a random number ri and then sends to one site and (i; ri) to the second site. Note that neither site willbe able to predict the xi. Due to the xor _ (operation), the input they see is indistinguishable from any uniformly generated random sequence. any data mining functionality can be evaluated privately without revealing any information other than the final result.While the architecture is not especially efficient, indeed not even necessarily very practical for large quantities of data, it does demonstrate a method of maintaining perfectprivacy while computing the required data mining function.

Secure computation and privacy-preserving data mining. There are two distinct problems that arise in the setting of privacy-preserving data mining. The rst is deciding which functions can be safely computed (\safely" meaning that the privacy of individuals is preserved). For example, is it safe to compute a decision tree on confidential medical data in a hospital, and publicize the resulting tree? This question is not the focus of this paper, but will be discussed brie°y in Section 5. For the most part, we will assume that the result of the data mining algorithm is either safe or deemed essential. Thus, the question becomes how to compute the results while minimizing the damage to privacy. For example, it is always possible to pool all of the data in one place and run the data mining algorithm on the pooled data. However, this is exactly what we don't want to do (hospitals are not allowed to hand their radata out, security agencies cannot the risk, and governments risk citizen outcry if they do). Thus, the question we address is how to compute the results without pooling the data ina way that reveals nothing but the internal results of the data mining computation. This question of privacy-preserving data mining is actually a special case of a long-studied problem in cryptography called secure multiparty computation. This problem deals with a setting where a set of parties with private inputs wishes to jointly compute some function of their inputs. Loosely speaking, this joint computation should have the property that the parties learn the correct output and nothing else, even if some of the parties maliciously collude to obtain more information. Clearly, a protocol that provides this guarantee can be used to solve privacy-preserving data mining problems of the type discussed

V. CONCLUSION

Privacy preserving has the more energy to increase the reach and benefits of data mining technology, Already we are taking the mixture of definitions. But these paper does not specify the some recommendations related privacy preserving in data mining technology. If we known the recommendations it is very easy to develop the standards and algorithms in privacy concern.

REFERENCES

1. P. Samarati and L. Sweeney. Protecting privacy whendisclosing information: k-anonymity and its enforcement through generalization and suppression. In Proceedings ofthe IEEE Symposium on Research in Security and Privacy,May 1998.
2. Y. Saygin, V. S. Verykios, and C. Clifton. Using unknowns to prevent discovery of association rules. SIGMOD Record,30(4):45–54, Dec. 2001.
3. L. Sweeney. Computational Disclosure Control: A Primer on Data Privacy Protection. PhD thesis, Massachusetts Instituteof Technology, 2001.
4. J. S. Vaidya and C. Clifton. Privacy preserving associationrule mining in vertically partitioned data. In TheEighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26 2002
5. A. C. Yao. How to generate and exchange secrets. In Proceedingsof the 27th IEEE Symposium on Foundations ofComputer Science, pages 162–167. IEEE, 1986.
6. D. Beaver, S. Micali and P. Rogaway. The Round Complexity of Secure Protocols. I22nd STOC, pages 503{513, 1990. 84
7. M. Bellare and S. Micali. Non-Interactive Oblivious Transfer and Applications. InCRYPTO'89, Springer-Verlag (LNCS 435), pages 547{557, 1989. 76
8. M. Ben-Or, S. Goldwasser and A. Wigderson. Completeness Theorems for Non- Cryptographic Fault-Tolerant Distributed Computation. In 20th STOC, pages 1{10, 1988. 66, 67, 84