Data Mining in Security, Surveillance, and Privacy Protection

Lisa Singh, Kun Liu

I. Data Mining in Security, Surveillance, and Privacy Protection

A. Introduction

While literature within the field of privacy-preserving data mining (PPDM) has been around for seven years, understanding the role of privacy in this context is still very much in its infancy. Most of the algorithms and approaches that have been introduced are very ad-hoc, computationally expensive, and do not support a formal theoretical approach similar to those that exist in security or databases. As a result, the applications of PPDM are quite rare in industry although there have been a plentiful publications in academia. This lack of standards and applications was echoed throughout the session. Therefore, as we consider issues and directions for the next generation of PPDM, we must preface this discussion with the need to establish a clear definition and theoretical framework for privacy. We must also bear in mind that this framework should be practically feasible in general as well as in the context of data mining applications.

We begin by describing the key contributions of each talk, identifying the recommendations made by each speaker. Please note that some of the talks in this session addressed areas outside of privacy. In this section, we only focus on the components of the talks related to privacy, security, surveillance and the like. We then integrate them and present a comprehensive set of recommendations, discussing each of them in more detail.

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

1) Talk overview: The speaker first gave a brief review of the history of PPDM. He pointed out that although PPDM 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).

2) 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?

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

1) Talk 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-mediated 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.

2) Recommendations:

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

Lisa Singh is with Georgetown University, email: singh@cs.georgetown.edu. Kun Liu is with IBM Almaden Research Center, email: kun@us.ibm.com.

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

1) Talk 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 ad hoc groups.

2) 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.

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

1) Talk overview: Much graph mining research to date focuses on simple network models containing a single node type and a 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.

2) 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.

F. Shashi Shekhar, Bhavani Thuraisingham, and Latifur Khan - Spatial and Spatio-temporal data mining challenges

1) Talk overview: The large number of geo-spatial data sets has given rise to spatial and spatial-temporal data mining. Applications include geo-spatial intelligence for security, surveillance of crime mappings for public safety, and containing the spread of infectious disease. Classical data mining approaches that assume independent, discrete transactions are not applicable since spatial data is highly auto-correlated and exhibit high degrees of heterogeneity. Further, existing technologies have inadequate models for dealing with richer temporal semantics and heterogeneity. The talk discussed a number of challenges for spatial data mining including: the validity of the independent assumption, the difficulty in capturing pattern continuity, the need to develop approaches for spatial anomaly and outlier detection. The second part of the talk focused on semantic integration of geospatial data. The speakers described a new approach using geospatial web services that integrates the OWL-S service ontology and the geospatial domain specific ontology to facilitate semantic matching services from multiple heterogeneous, independent data sources. They illustrated this approach using data created from ASTER, a thermal emission and reflection radiometer. While the approach is promising, a number of challenges still need addressing, including pixel and semantic merging of neighborhood regions, handling irregular shapes, scalability, and developing security policies for geo-spatial data. Privacy and security have not been sufficiently explored in the context of geo-spatial data. What is an individual's right to privacy? For example, if Google maps can capture an individual's residence, how can the individual maintain privacy?

2) Recommendations:

- Develop scalable approaches for detecting multiple spatial anomalies and for handling uncertain, heterogeneous spatial data.
- Develop richer temporal semantics for geo-spatial data. Current models are inadequate.
- Develop test data sets that can be used to evaluate different methods for spatial-temporal data mining.
- Define privacy for geospatial data and consider implications of public source geo-spatial data on individual privacy.

G. Michael Berry - Automating the Detection of Anomalies and Trends from Text

1) Talk overview: Nonnegative matrix factorization (NNMF) has been widely used to approximate high dimensional data comprised of nonnegative components. This talk presented a NNMF algorithm for detecting anomalies and trends from unstructured text documents. By preserving nonnegativity, NNMF extracts concepts and topics from the document and enables a non-subtractive combination of parts to form a whole. More specifically, the algorithm parses the documents and produces a reduced-rank representation of an entire document space. The resulting feature and coefficient matrix factors are then used to cluster the documents. The speaker demonstrated the performance of the algorithm by using data from the Aviation Safety Reporting System (ASRS). The results show that anomalies of the training documents can be directly mapped to NNMF-generated feature vectors. Dominant features of test documents can then be used to generate anomaly relevance scores for effective anomalies detection. The speaker also explored the challenges of using Multiplicative Method (MM) for solving NNMF.

2) Recommendations:

- Develop scalable, robust and incremental algorithms for solving NNMF.
- Identify ways to interpret the features generated by NNMF. How sparse (or smooth) should factors (W, H) be in order to produce as many interpretable features as possible?
- Study NNMF on multimodal data. Can we build a nonnegative feature space from objects in both images and text? Are there opportunities for multimodal document similarity?

H. Joe Kielman and Ted Senator- Panel on Future Research Challenges and Needed Resources for "Data Mining in Security, Surveillance, and Privacy Protection"

1) Talk overview: The first talk was given by Dr. Kielman from Department of Homeland Security. Dr. Kielman expressed his concerns about public surveillance – a huge amount of data about individuals are constantly being collected through surveillance systems by governmental and private organizations. Using this information effectively is an important issue that has not been carefully investigated. He pointed out that it might be interesting to develop privacy protection techniques that work in real time. His talk also covered other issues such as privacy of DNA figure prints. Dr. Kielman particularly suggested to develop systems that do not use real private data but simulated synthetic data, which is a kind of research they are interested to fund.

The following talk was presented by Dr. Senator from SAIC. Dr. Senator first discussed the definition of "data mining" from both technical and political perspectives. Next, he defined his notion of privacy, which is the ability to prevent linkage of identity to information. Then, he offered some recommendations for doing data mining responsibly: 1) consider privacy implications before beginning the project; 2) be completely transparent regarding the purpose of data collection, how will the data be used, who will have access to it, how it will be secured, where and for how long is the data retained, can individuals access and correct their personal information, etc. He emphasized the fact that technology is only part of the solution for privacy protection while data, processes, policies, authorities, and laws are at least as important and difficult. Finally, he suggested some research problems and needed recourses for the next generation of privacy-preserving data mining.

2) Recommendations:

- Develop techniques to protect genomic information.
- Develop scalable, real time privacy protection techniques.
- Develop synthetic data generation techniques for privacy protection.
- Develop identity-free pattern discovery techniques.
- Develop network/graph anonymization algorithms.
- Develop provably auditable data mining systems.
- Develop privacy enforcement mechanisms and formalize privacy policies.
- Develop relationship-preserving anonymization algorithms.
- To achieve the goal, we need privacy officers who understand technology as well as scientists, managers, and users who understand privacy.

I. Overall Recommendations

Pioneered by Agrawal & Srikant and Lindell & Pinkas' work from 2000, there has been an explosive number of publications in privacy-preserving data mining. Many techniques have been proposed, questioned, and improved. However, compared with the active and fruitful research in academia, applications of privacy-preserving data mining for real-life problems are quite rare. Without practice, it is feared that research in privacy-preserving data mining will stagnate. Furthermore, lack of practice may hint to serious problems with the underlying theories/concepts of privacy-preserving data mining. Identifying and rectifying these problems must be a top priority for advancing the field.

This session served as a forum for researchers, scientists, and engineers to discuss the challenges and opportunities of data mining in privacy protection, surveillance and security. We integrate the contributions from all the speakers and present a comprehensive set of recommendations for the next generation of research.

- Develop a formal theoretical framework for privacy. What is privacy? What is privacy-preserving data mining? What problems are we trying to solve and are the current state-of-the-art approaches reasonable solutions? How does privacy different from security, anonymity, cryptography, etc.?
- Develop privacy models and techniques for specific domains. It is clear that privacy is a domain dependent concept: homeland security, healthcare, business secrecy, entertainment, and ubiquitous computing are each different and pose different privacy requirements. Using the general, formally defined theory of privacy developed in recommendation 1, we should extend the models, evaluation techniques, and privacy metrics for these different domains.
- Develop approaches for complex data. Most data is not simple and independent, e.g. temporal spatial, social networks and graphs, and text. Therefore, for privacy to be useful, it must be a viable option for domains containing complex, dynamically changing data.
- Develop metrics that adequately quantify levels of privacy in different types of data. How much inferences exists in a real world data set prior to use of privacy-preserving techniques? What are 'acceptable' levels of privacy, both from a policy perspective and from a computational one?
- Understand economic and legal aspects of privacy protection. Privacy is no less a societal and economical concept than it is a technological challenge. However, the majority of work on privacy-preserving data mining does not focus on these aspects. Future research directions could include: 1) the economics of privacy; 2) modeling of privacy legislation and automated proofs of adherence; and 3) privacy and utility trade-off.
- Develop practical performance-aware PPDM techniques. Some of the technological impediments to privacy are rooted in the performance of the algorithms. Future work should pay attention to this issue. Research includes efficiency improvements to known algorithms, scalable privacy models, etc.
- Create a benchmark data set repository for data that needs to be privatized. This would let researchers compare algorithms on known data sets to more clearly understand the differences among various approaches.

Privacy-preserving data mining is a fruitful area that has been slow to gain steam. As more researchers, engineers and legal experts delve into this area, standards and theory will begin to take shape. As these are established, the next generation of PPDM will be a fertile ground for all concerned with privacy implications in society.