PinKDD'08: Privacy, Security, and Trust in KDD Post-Workshop Report

Francesco Bonchi
Yahoo! Research,
Barcelona, Spain
bonchi@yahoo-inc.com

Elena Ferrari
Computer Science and Communication Dept.
University of Insubria
Varese, Italy
elena.ferrari@uninsubria.it

Wei Jiang
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO, USA
wjiang@mst.edu

Bradley Malin
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN, USA
b.malin@vanderbilt.edu

ABSTRACT

This report summarizes the events of the 2^{nd} International Workshop on Privacy, Security, and Trust in KDD, at the 14^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The workshop was held on August 24, 2008 in Las Vegas, Nevada and brought together computer scientists working on how data protection issues factor into the context of data mining.

1. INTRODUCTION

The field of computer science is evolving to incorporate intrinsically complex social and organizational environments. There is an ever-increasing demand for the incorporation of new technologies to gather data on people for a variety of worthwhile endeavors. Nowhere is this movement more apparent, and the influence of data mining professionals more critical, than in the often debated arenas of privacy, security, and trust. The increased collection and sharing of personal information for data mining endeavors raises complex societal issues regarding data management and civil liberties. Ensuring data protection is essential for the provision of electronic and knowledge-based services in modern e-business, e-commerce, e-government, and e-health environments. If insufficiently addressed, mishandling of the data will harm both the corresponding persons and data holders, leading to a loss in the public's trust and legal action. To prevent such problems, computer scientists are developing novel techniques, as well as recasting cryptographic tools from the security domain, to simultaneously ensure privacy while facilitating data mining projects.

To inject privacy and trust into security and surveillance data mining projects, it is necessary to facilitate a dialogue between researchers and practitioners in the associated communities. Last year, we organized the first instantiation of this workshop (PinKDD'07) to introduce researchers from disparate environments, including business, security, and theory to learn about the concerns and potential solutions regarding their challenges. This year's workshop continued the integration of researchers investigating privacy, security, and issues of trust within a data mining framework.

2. WORKSHOP SUMMARY

PinKDD'08 received many high-quality research paper submissions, each of which was reviewed by a minimum of three members of the program committee. In all, six papers were selected for presentation at the workshop and inclusion in the workshop's post-proceedings. At the workshop, the research presentations were grouped into two themes 1) *"privacy protection through anonymity"* and 2) *"attack and detection"*. In addition, the workshop included a keynote talk and an invited session on Geospatial Privacy Protection.

2.1 Keynote Talk

The workshop began with a keynote talk delivered by Prof. Thuraisingham of the University of Texas at Dallas, entitled *"Privacy, Security and Trust for Data Mining "*. Prof. Thuraisingham provided an overview of the relationships of data mining to privacy, security and trust. Prof. Thuraisingham proceeded to summarize various solutions to the related problems. This talk concluded by emphasizing the importance that policy makers, legal analysts, technologists, security experts and privacy advocates work together to develop flexible and practical solutions for data mining.

2.2 Session One: Protecting Anonymity

The first research session focused on papers that addressed models and algorithms for the protection of personal privacy through data anonymization and distortion techniques.

The first paper received the best paper award of PinKDD'08, sponsored by Yahoo! Research. This work investigated privacy protection in social networking research. Massive quantities of information provided by the users social networking domains is collected, stored, and may be used for various purposes. Technology and policy researchers alike have pointed out various privacy breaches that can arise from sharing of social network data. Though efforts have been made to protect such data from unauthorized disclosure,

most data privacy research has focused on traditional data management models such as relational forms. Yet, unlike relational models, social networks contain relationships among individual entities. Campan et al. [1] developed a greedy algorithm for anonymizing a social network and introduced a structural measure to quantify the amount of information lost due to edge generalization in the anonymization process.

Data distortion is a common technique in the field of privacy preserving data mining. The premise of this technique is original data is hidden through the addition of noise from known distributions. Data mining models can then be built from the distorted data. Under this paradigm, Rachlin et al. [5] introduced a definition of guessing anonymity, which captures the difficulty of achieving an external linking attack. Based on this definition, methods were proposed to select the appropriate perturbation parameters for satisfying a desired privacy protection.

2.3 Session Two: Detecting Malicious Activity

The second research session of the workshop was dedicated to papers that focused on frameworks and algorithms to detect malicious activities related to web-based applications, such as the detectiion of anomalies in web services and the development of attack models in recommender systems, and the detection of unknown malicious code.

In the first paper, Li et al. [3] presented an efficient network anomaly detection method based on the k-nearest neighbor data mining algorithm. Their approach integrated objective functions and anomaly impact metrics from the end users' perspective to ensure the robustness of the web server anomaly detection mechanism. Li introduced a generic algorithm based on an instance selection mechanism to improve the real-time detection performance.

In the second paper, Castellano et al. [2] presented a flexible framework, consisting of e-services and data mining components, for intrusion detection. The goal of the framework is to provide security managers with modular and flexible functionalities to resolve intrusion problems. In particular, Castellano addressed a class of possible solutions based on knowledge discovery using data mining and web mining technologies. This work demonstrated how intrusion detection systems can be built using the proposed framework.

Malicious users can influence online recommendation systems by providing biased data. Such attacks generally lead to the erosion of user trust in the objectivity and accuracy of the system. In the third paper, Ray et al. [6] presented new attack strategies that explore the importance of the targeted items and "filler" items. Unlike previous approaches, their strategies were built specifically for user-based and item-based collaborative filtering systems, with a particular focus on the intelligent selection of filler items. Empirical results demonstrated that the strategies may be effective against both types of filtering environments.

The growth of network usage has motivated the creation of emerging malicious code for various purposes. Although signature-based anti-viruses methods are accurate, they cannot detect newly created malicious code. In the final paper, Moskovitch et al. [4] presented a methodology to detect unknown malicious code based on text categorization concepts. Through an active-learning framework, which enables the selection of unknown files for fast acquisition, the proposed methodology can be further improved to achieve a more accurate and efficient acquisition of unknown malicious files.

2.4 Invited Session: Geospatial Privacy

The workshop concluded with an invited session on Geospatial Privacy Protection featuring two talks by Peter Christen and Franco Turini, followed by an open discussion. Prof. Christen's talk, "Geocode Matching and Privacy Preservation", started with a brief introduction to privacy preserving data matching and record linkage. Using a variety of real-world scenarios, the talk illustrated the multitude of privacy and confidentiality inherent in geocode matching. Next, Prof. Turini gave a talk entitled "Mobility, Data Mining and Privacy". The talk highlighted the results of a European-wide research project called GeoPKDD, EU-funded Geographic Privacy-Aware Knowledge Discovery and Delivery (http://www.geopkdd.eu). During the delivery, particular attention was placed upon privacy-aware spatio-temporal data mining from mobility data generated by wireless networks, mobile technologies, and ubiquitous computing.

3. ACKNOWLEDGMENTS

We thank the authors of all submitted papers, the invited speakers, and all attendees for contributing to the success of the workshop. We would also like to express our gratitude to the members of the Program Committee for their vigilant and timely reviews, namely: Maurizio Atzori, Elisa Bertino, Barbara Carminati, Peter Christen, Chris Clifton, Josep Domingo-Ferrer, Tyrone Grandison, Dawn Jutla, Murat Kantarcioglu, Ashwin Machanavajjhala, Stan Matwin, Taneli Mielikinen, Yucel Saygin, Kian-Lee Tan, Bhavani Thuraisingham, Vicenç Torra, Vassilios Verykios, Ke Wang, Rebecca Wright and Jeffrey Yu. Finally, we thank the sponsors: The UNESCO Chair in Data Privacy, Yahoo! Research, IBM Research, and GeoPKDD, a project in the EU Future Emerging Technologies program.

4. REFERENCES

[1] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.

[2] M. Castellano, G. Mastronardi, G. Decataldo, L. Pisciotta, and G. Tarricone. Composing miners to develop an intrusion detection. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.

[3] Y. Li, L. Guo, and Z.-H. Tian. Web server anomaly detection via lightweight tcm-knn data mining scheme. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.

[4] R. Moskovitch, N. Nissim, and Y. Elovici. Acquisition of malicious code using active learning. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.

[5] Y. Rachlin, K. Probst, and R. Ghani. Maximizing privacy under data distortion constraints in noise perturbation methods. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.

[6] S. Ray and A. Mahanti. Strategies for effective shilling attacks against recommender systems. In *Proc. 2nd Int'l Workshop on Privacy, Security, & Trust in KDD*, Las Vegas, NV, USA, 2008.