

PinKDD'07: Privacy, Security, and Trust in KDD Post-Workshop Report

Francesco Bonchi
Pisa KDD Laboratory,
ISTI-C.N.R.
Pisa, Italy

francesco.bonchi@isti.cnr.it

Bradley Malin
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN, USA

b.malin@vanderbilt.edu

Elena Ferrari
Computer Science and Communication Dept.
University of Insubria
Varese, Italy

elena.ferrari@uninsubria.it

Yücel Saygin
Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey

ysaygin@sabanciuniv.edu

ABSTRACT

In this report, we summarize the events of the First International Workshop on Privacy, Security, and Trust in KDD (PinKDD), which was held in conjunction with the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The workshop convened on August 12, 2007 in San Jose, California and brought together researchers, as well as practitioners, working on how privacy, security, and trust can be resolved or modeled within a data mining framework.

1. INTRODUCTION

Vast amounts of data are collected by service providers, system administrators, and are available in public information systems. Data mining technologies provide an ideal framework to assist in the analysis of such collections for computer security and surveillance-related endeavors. For instance, system administrators can apply data mining to summarize activity patterns in access logs so that potential malicious incidents can be further investigated. Beyond computer security, data mining technology supports intelligence gathering and reporting for homeland security. For years, and most-recently fueled by events such as September 11, 2001, government agencies have focused on developing and applying data mining technologies to monitor terroristic behaviors in public and private data collections.

The application of data mining to person-specific data raises serious concerns regarding data confidentiality and citizens privacy rights. These concerns have promulgated the adoption of various legislation and policy controls. In 2006, the European Union passed a data-retention directive that requires all telephone and Internet service providers to store data on their consumers for up to two years to assist in the prevention of terrorism and organized crime. [4] Similar data-retention regulation proposals are under heated debate in the United States Congress. However, the debate often

focuses on ethical or policy aspects of the problem, such that resolutions have polarized consequences; e.g. an organization can either share data for data mining purposes or it can not. Fortunately, computer scientists, and data mining researchers in particular, have recognized that technology can be constructed to support more-flexible solutions. Computer scientists are developing technologies that enable data mining goals without sacrificing the privacy and security of the individuals to whom the data corresponds.

To inject privacy into security and surveillance data mining projects, it is necessary to understand the goals of the latter. To further this exchange and highlight advances in research, we organized the First International Workshop on Privacy, Security, and Trust in KDD (PinKDD).

2. WORKSHOP SUMMARY

The PinKDD workshop attracted considerable attention from the research community, as well as support from industrial organizations and academic institutions. The workshop received many high-quality research paper submissions, each of which was reviewed by a minimum of three members of the program and organizing committee. In all, eight papers were selected for presentation at the workshop and inclusion in the workshop's post-proceedings. The papers represented the diversity of data mining research issues in privacy, security, and trust.

At the workshop, the research presentations were grouped into two themes 1) "*privacy preserving data mining with multiple, or distributed, datasets*" and 2) "*anonymity, web, and graph privacy*". In addition to two research sessions, the workshop highlights included a keynote talk and a spirited panel discussion on privacy issues in weblogs.

2.1 Keynote Talk

The workshop began with a keynote talk, "*An Ad Omnia Approach to Defining and Achieving Private Data Analysis*", which was delivered by Dr. Cynthia Dwork of Microsoft Research. In this talk, Dr. Dwork provided formal definitions and models of privacy protection in databases. Based on

formal models of the system, Dr. Dwork, proceeded to summarize research that characterizes the theoretical limits of privacy protection that can be achieved in databases. This talk concluded by highlighting recent work conducted by Dr. Dwork and colleagues on the ability, or lack thereof, to protect privacy in highly complex datasets, such as social networks.

2.2 First Session: Privacy Preserving Mining with Multiple/Distributed Datasets

The first research session of the workshop was dedicated to papers that concentrated on algorithms and data structures to achieve privacy in distributed datasets. Privacy preserving data mining in distributed settings has become a crucial topic, due in part to the rapid growth of ubiquitous computing environments. When data is distributed across a set of parties, the involved parties have the opportunity to perform data mining on a larger quantity of information and thus learn more robust and accurate rules and models. A common goal in privacy preserving distributed data mining is to merge models learned from local datasets to construct a global model without revealing sensitive local information. In the first presentation, Sharkey et al. [8] tackled this problem and proposed a method for knowledge sharing among multiple parties. Specifically, they focused on how to learn decision tree models. The main contribution of their approach was that it reduced the computational and communication complexity that is inherent to earlier data exchange protocols for distributed privacy preserving data mining.

Model sharing needs to be supported by a mechanism to tune the model or to select the best model among the possible alternatives. The work presented by Yang et al. [10] proposed a method for multiple parties to perform model selection in a privacy preserving fashion, without revealing any data to each other. They consider the popular cross-validation method for model selection and demonstrate how to mine vertically partitioned data in a secure setting for two parties.

When data owners lack expertise or computational power, a central authority is needed to perform data mining tasks. One way this architecture can be leveraged is for each data owners to perturb their records before submitting data to the central repository. The goal of perturbation is to prevent the revelation of the original values of any particular record, while retaining enough information in the data so that it remains useful for data mining purposes. Tan and Ng [9] provided the third presentation of this session and described a new privacy-preserving distributed data sanitization algorithm. In this algorithm, they build a classifier by independently randomizing the private data at each site before the data is fed into the central store.

Most prior distributed privacy preserving data mining algorithms assume a “semi-honest” model, which is a very strong assumption that is inappropriate for certain settings. Ahmad and Khokhar [1] addressed this issue in the third presentation in which they described an efficient privacy preserving clustering algorithm on horizontally partitioned data under the malicious adversary model. In this setting, the authors presented a protocol that does not require a central authority.

2.3 Second Session: Anonymity, Web and Graph Privacy

The availability of detailed person-specific data sets for data mining and personalization research in domains such as web usage/search and healthcare is limited due to privacy concerns. De-identification, or the removal of explicit identifiers, has often been used as a data preprocessing step to protect personal identities before data publication. However, it has been shown that the combination of a subset of the attributes of the de-identified data, called a quasi-identifier, can often be used to link seemingly anonymous data to named individuals in publicly available resources. A recent example of data “re-identification” occurred in August 2006 when several AOL researchers disclosed the de-identified web logs of approximately 20,000 AOL users. Each user’s identity was replaced with a consistent pseudonym, but various re-identifications were conducted and highly publicized by journalists that linked search terms, such as Social Security Number, addresses, or personal names, retained in the logs to the individuals that issued the queries. [2; 5]

One of the reasons why re-identifications occur is that data owners do not recognize which features constitute a quasi-identifier prior to disclosing data collections. In the first presentation of the second session, Lodha and Thomas [6] provided a formal characterization and a technique by which data owners can discover quasi-identifiers. They applied the characterization to derive a probabilistic notion of anonymity.

In the second presentation, Chaytor [3] tackled the problem of generating a k -anonymous dataset through suppression. A dataset is said to be k -anonymous when the quasi-identifying values of each record is equivalent to $k-1$ other records. This is a simple concept, but one in which finding the minimal amount of information to suppress is known to be an NP-hard problem. Chaytor proposed a new genetic algorithm that represents column orderings as permutations and adopts an ordered greedy approach that had not been considered before for anonymization. The approach was experimentally validated to be more adept at minimizing information loss than prior attempts.

The third presentation, by Zheleva and Getoor in [11], addressed the problem of link re-identification. In this problem, the goal is to protect sensitive relationships in de-identified graph data, such as the relationships derived from a social networking website. In this work, they proposed five different privacy preservation strategies, each with different tradeoffs between data utility and privacy preservation.

The best paper award was presented to Poblete et al. [7] for their paper entitled “Website Privacy Preservation for Query Log Publishing”. In the fourth presentation of this session, the authors introduced a new privacy concern, *website privacy* and defined the possible adversaries that could reveal proprietary website information via information in web search query logs. To overcome such website privacy leaks, the authors proposed anonymization techniques for various types of attacks that could be attempted by an adversary.

2.4 Panel: Privacy Challenges and Opportunities for Sharing and Mining Weblogs

The continuing growth of online search engines, such as Google, Yahoo, and Microsoft, generates incredible quantities of detailed personal information. Every day, people looking for information on the World Wide Web submit millions of search queries. This information is stored and, in many instances, is mapped to an individual's prior searches. Search queries are only one type of information collected by websites through user interaction and as the amount, and diversity, of person-specific data stored at websites grows, so too do the data mining opportunities for service personalization. However, the events of the AOL search log re-identifications, and the work presented by Poblete et al. [7], provided clear illustrations of how weak protections can lead to significant breaches of user's expected, as well as promised, privacy. The goal of this panel was to gather experts on data mining, weblogs, and privacy to discuss the current issues and potential directions for research in the area. The panel consisted of Dr. Ricardo Baeza-Yates (Yahoo Research), Dr. Cynthia Dwork (Microsoft Research), Dr. Lise Getoor (University of Maryland, College Park), and Dr. David Jensen (University of Massachusetts Amherst).

3. CONCLUSIONS AND FUTURE DIRECTIONS

Privacy, security, and trust in data mining are related issues that have captured the attention of many researchers, administrators, and legislators. Consequently, data mining for improved security, and the study of data mining side-effects on privacy, has rapidly become a hot and lively research area. The issues are rooted in the real-world and concern academia, industry, government, as well as society in general. The issues are global and many governments are struggling to set national, and international, policies on privacy and security for data mining endeavors. The PinKDD workshop demonstrated that privacy-aware data mining technologies can be developed, but also that there are limitations to existing computational models and privacy policies. The analysis of the security and privacy aspects of data mining have begun, and the research and debate on display at PinKDD illustrated that the topics is moving towards maturity and main stream acceptance. We commend all of participants and look forward to continuing advances in the field!

4. ACKNOWLEDGMENTS

We would like to thank the authors of all submitted papers, the invited speaker, the panelists, 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, Roberto Bayardo, Barbara Carminati, Peter Christen, Josep Domingo-Ferrer, Wenliang (Kevin) Du, Tyrone Grandison, Satoshi Hada, Dawn Jutla, Murat Kantarcioglu, Hillol Kargupta, Stan Matwin, Ilya Mironov, Taneli Mielikinen, David Skillicorn, Kian-Lee Tan, Bhavani Thuraisingham, Vicenç Torra, Vassilios Verykios, Ke Wang, and Jeffrey Yu.

Last, but not least, we acknowledge the fundamental support of our sponsors: The UNESCO Chair in Data Privacy,

Agnik LLC, the *Context, Content and Community* team from Nokia Research Center Palo Alto, KDubiq (Knowledge Discovery in Ubiquitous Environments) European coordination action, and GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery) a European project within the Future Emerging Technologies program of FP6-IST.

5. REFERENCES

- [1] W. Ahmad and A. Khokhar. Phoenix: Privacy preserving biclustering on horizontally partitioned data amid malicious adversaries. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [2] M. Barbaro and T. Zeller. A face is exposed for aol searcher no. 4417749. *New York Times*, August 9, 2006.
- [3] R. Chaytor. A better problem representation for k -anonymity. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [4] European Union. Directive 2006/24/ec of the european parliament and of the council of 15 march 2006. *Official Journal of the European Union*, L 105, 54, April 13, 2006.
- [5] S. Hansell. Aol removes search data on group of web users. *New York Times*, August 8, 2006.
- [6] S. Lodha and D. Thomas. Probabilistic anonymity. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [7] B. Poblete, M. Spiliopoulou, and R. Baeza-Yates. Website privacy preservation for query log publishing. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [8] P. Sharkey, H. Tian, W. Zhang, and S. Xu. Privacy-preserving data mining through knowledge model sharing. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [9] V. Tan and S.-K. Ng. Privacy-preserving sharing of horizontally-distributed private data for constructing accurate classifiers. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [10] Z. Yang, S. Zhong, and R. Wright. Towards privacy preserving model selection. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.
- [11] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *Proceedings of the First International Workshop on on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, California, August 2007.