Preface

Fourth International Workshop on Privacy Aspects of Data Mining (PADM 2013)

Introduction

Machine learning and data mining algorithms have penetrated our everyday lives and play a central role in many application domains, including social networks, healthcare, location-based systems, and advertising. At the same time, 90% of today's data have been produced in the last two years! These data come from social networking sites, mobile phone applications, electronic medical record systems, e-commerce sites, and open data portals. The analysis of this wealth of data can lead to valuable insights that will benefit data recipients and the society at large, but may also lead to serious privacy breaches, unless privacy-enhancing technologies are in place.

To address the increasing privacy concerns of both individuals and organizations about the use of their data, the domain of privacy-preserving data mining was brought into existence more than a decade ago. Research in this area has led to a wide range of privacy methods proposed by industry and academia, such as the development of differential privacy and utility-preserving anonymity. However, emerging applications pose new research challenges for privacy-preserving methods, which are related to the type and intended use of the data, while privacy should be embedded throughout the lifecycle of the data, from data collection to data sharing and use. In fact, there are several interesting issues that require further investigation, both from a theoretical and a practical perspective.

First, there is an urgent need to address the challenges of preserving privacy in emerging applications, such as (location-based) social networks, healthcare, cloud computing, e-commerce and location-based systems. Protecting data in these applications is particularly challenging, because they involve large volumes of complex, heterogeneous, and dynamic data. These data must be protected from several attacks that may lead to disclosure of private or sensitive information and discrimination, using scalable methods.

Second, ensuring that privacy-protected data remain useful in intended applications, such as building accurate data mining models or enabling complex analytic tasks to be performed, is essential. This calls for a careful study of data utility requirements, and the development of advanced algorithms to enforce these utility requirements while protecting data privacy. Meanwhile, benchmarks for evaluating data utility and large-scale case studies, which will support the reliable and meaningful comparison of competing methods, need to be developed.

Third, privacy-preserving data integration and linkage still remain very challenging tasks that call for new approaches. This is because there is a large number of data providers and recipients, with different and possibly competing privacy and utility requirements, who need to co-operate in order to produce privacy-protected data or collaborate to compute a common function in a privacy-preserving way.
Workshop Summary
The 4th International Workshop on Privacy Aspects of Data Mining (PADM 2013) was a full-day workshop held in conjunction with the IEEE ICDM conference in Dallas, Texas, on December 7, 2013. Its aim was to facilitate the exchange of ideas among researchers, to highlight the most recent advances in the area of privacy-preserving data mining, and to encourage principled research that develops methodologies to address open and challenging privacy issues.

The workshop attracted significant attention from the research community, leading to several high-quality research paper submissions, each of which was reviewed by a minimum of three members of the program and organizing committee. In all, seven papers were selected for presentation at the workshop and inclusion in the workshop's post-proceedings, including two papers that were transferred from the ICDM conference. Among the accepted papers, five manuscripts were selected to be invited to special sections that were organized after the workshop in the Journal of Privacy and Confidentiality (http://repository.cmu.edu/jpc/) and the Transactions on Data Privacy (http://www.tdp.cat), and are scheduled to appear in 2014.

Research Sessions
At the workshop, the research presentations were grouped into three sessions:
(i) privacy models and algorithms for emerging applications,
(ii) theory and methodologies for discrimination-aware data mining, and
(iii) techniques for privacy-preserving data mining.

The first research session of the workshop was dedicated to privacy for emerging applications. Three papers were presented in this session that considered how to protect data, such as user trajectories, social networks, or aggregate time series, intended for different applications.

In the second research session, two papers that proposed new theory and methodologies for discrimination-aware data mining were presented. These papers examined interesting theoretical connections between anti-discrimination and anonymization, and they also proposed methods for avoiding discrimination.

The third session was devoted to privacy-preserving data mining. The authors of the two accepted papers proposed approaches to outsource the computations of a clustering algorithm in a privacy-preserving way and to construct an accurate, differentially-private classifier.

Keynote Talks
In addition to the three research sessions, the workshop highlights included two invited keynote talks that were delivered by Prof. Li Xiong of the Emory University in Atlanta, Georgia, and by Prof. Murat Kantarcioglu of the University of Texas at Dallas.

The first keynote talk was on adaptive differentially-private data release for data sharing and data mining. In this talk, Prof. Xiong presented an overview of the Adaptive Differentially Private Data Release (ADP) project, which aims at offering a suite of data-driven and adaptive techniques for releasing differentially private data by exploiting the characteristics of the underlying data. As part of this project, various types
of data have been considered for release, including relational, transactional, sequential, and time series data, with the purpose of supporting different analytic tasks, such as classification and frequent pattern mining. The talk concluded with a discussion on the important research challenges and open questions of applying differential privacy to facilitate general-purpose data sharing.

The second keynote talk was delivered by Prof. Kantarcioglu and was on incentive-compatible privacy-preserving distributed data mining. In this keynote talk, Prof. Kantarcioglu presented an overview of past privacy-preserving distributed data mining work, involving cryptographic protocols that could be implemented by using a few basic secure blocks (sub-protocols), and then discussed two incentive-compatible privacy-preserving distributed data mining protocols that were recently proposed. The latter protocols support the integration and mining of large amounts of privacy-sensitive distributed data, without violating privacy and security constraints. A major issue that these protocols try to address is that of data truthfulness, i.e., how to ensure that the players participating to a secure cryptographic protocol will provide true data for the performed calculation. In order to achieve that, incentive-compatible protocols audit the process with a high enough frequency and stiff enough penalty to prohibit players from lying about their data. As a case study for the usefulness of these protocols, Prof. Kantarcioglu presented their use in the biomedical domain for enabling the privacy-preserving distributed mining of genomic sequences.

Organizers
Aris Gkoulalas-Divanis, IBM Research, Ireland
Grigorios Loukides, Cardiff University, UK
Charu Aggarwal, IBM T. J. Watson Research Center, USA

Steering Committee
Raghav Bhaskar, Microsoft Research, India
Aris Gkoulalas-Divanis, IBM Research, Ireland
Dan Kifer, Pennsylvania State University, USA
Srivatsan Laxman, Microsoft Research, India
Lisa Singh, Georgetown University, USA
Xintao Wu, University of North Carolina at Charlotte, USA

Program Committee
Francesco Bonchi, Yahoo! Research Barcelona, Spain
Maria Luisa Damiani, University of Milan, Italy
Josep Domingo-Ferrer, Universitat Rovira i Virgili, Catalonia
Benjamin Fung, Concordia University, Canada
Wei Jiang, Missouri University of Science and Technology, USA
Murat Kantarcioglu, University of Texas at Dallas, USA
Guenter Karjoth, IBM Research-Zurich, Switzerland
Panagiotis Karras, Rutgers University, USA
Bradley Malin, Vanderbilt University, USA
Kun Liu, Yahoo! Labs, California, USA
Mehmet Sayal, Hewlett Packard, USA
Jianhua Shao, Cardiff University, UK
Lisa Singh, Georgetown University, USA
Spiros Skiadopoulos, University of Peloponnese, Greece
Aleksandra Slavkovic, Penn State University, USA
Adam Smith, Penn State University, USA
Manolis Terrovitis, Institute for the Management of Information Systems, Greece
Philip S. Yu, University of Illinois at Chicago, USA
Ke Wang, Simon Fraser University, Canada