Review of the book

"Privacy-Aware Knowledge Discovery"

Edited by Francesco Bonchi and Elena Ferrari
CRC Press
2011

ISBN: 978-1-4398-0365-3

Eric Diehl
Technicolor Security Office, Technicolor

2013-09-03

What the book is about

Data mining or knowledge discovery is one of the current hot topics. There is no doubt that its benefits will be decisive for academics, governments, companies and end users. Nevertheless, there are serious related privacy issues. This book presents some of the related privacy issues and their potential solutions. It explores several domains: data such as traces and streams, data related to movement and localization, time series, biomedical data, data related to web usage and social networks. The four first chapters present the techniques generally used to anonymize data. The remainder of the book explores the mentioned domains.

What the book is like

Chapter 1: *Anonymity Technologies for Privacy-Preserving Data Publishing and Mining*

This first chapter introduces the problem of anonymity and drafts the tradeoff between utility and privacy. Then, it introduces the two basic concepts used to anonymize a dataset: by randomization, and by indistinguishability, through masking, or synthetic techniques that create news data. A section is dedicated to privacy in regulations. Unfortunately, this section is too short to have a real interest. A point that I appreciated in this chapter is the particular focus on the need of an attacker model.

Chapter 2: *Privacy Preservation in the Publication of Sparse Multidimensional Data*

This chapter details more the key concept of k-anonymity and l-diversity using examples. The focus is on small dimension data. Other chapters will tackle the issue of large dimensional data, for instance chapter 13.

Chapter 3: *Knowledge Hiding in Emerging Application Domains*

The purpose of knowledge hiding is to conceal patterns which may be sensitive inside a database to be published. Knowledge is classified following two dimensions: sensitivity and derivability. Derivability is the ability to gain this knowledge through data mining the database. Thus, the generic knowledge hiding problem requires building a sanitized database such as no sensitive data can be derived while the quality of the sanitized database is similar to that of the genuine database.

The chapter studies concrete examples. The first one is to hide frequent item sets as found in the market based analysis. The other classes of problems are association rule hiding, sequential pattern hiding and classification rule hiding. For each class, the authors give an overview of the currently proposed solutions. The first category is more extensively described than the other ones.

Chapter 4: *Condensation-Based Methods in Emerging Application Domains*

Condensation is one method used to preserve privacy. Condensation regroups data within larger groups like rain drops. The output groups still preserve some statistical properties of the original dataset (for instance first and second order statistics). The chapter briefly presents the techniques applied to sequence data types such as tabular data, strings, sequential pattern, and trajectories.

Chapter 5: *Catch, Clean, and Release: A Survey of Obstacles and Opportunities for Network Trace Sanitization*

This chapter introduces the notion of sanitization. In addition to preserving the privacy of users (anonymization), sanitization protects also secrecy of other characteristics. In this case, it is the details of the traced network. The authors cite four large initiatives of network data sharing: CAIDA, CRAWDAD, PREDICTN and DSHIELD. They propose a taxonomy of sanitization techniques and focus on the prefix-preserving IP address issues. After that, they give an overview of some desanitization techniques. The conclusion is that currently desanitization always wins.

Chapter 6: *Output Privacy in Stream Mining*

The authors introduce the problem by highlighting the attacks via inference. The research of rare patterns may disclose some private information. Then they present two approaches to solve this issue: detecting and removing strategy with a list of solutions, and the proactive approaches (with exclusively the authors' work cited). The authors claim that the second approach is better than the first one. In the remainder of the chapter, they describe their solution: BUTTERFLY.

Chapter 7: *Privacy Issues in Spatio-Temporal Data Mining*

This chapter explores privacy within spatio-temporal data mining, such as trajectories. It first illustrates the risks. Then, the authors present some strategies such as suppressing samples so that to remove vulnerable item sets, a coarsening strategy that attempts to reach k-anonymity by coclustering k similar trajectories. Then, they explore some use cases. For instance, in secure multiparty computation, different parties collectively mine their dataset while not disclosing their own datasets, or hiding sequential patterns.

Chapter 8: *Probabilistic Grid-Based Approaches for Privacy-Preserving Data Mining on Moving Object Trajectories*

This chapter studies how to protect location privacy. Indeed, with traces of all users' movement and inferring with other information, it is possible to guess the identity of the user (using her home address and work address guessed from the most frequently visited locations). They provide a short overview of the state of the art, especially for location based services (LSB) and spatio-temporal data mining. Then, the chapter focuses on spatio-temporal anonymity of trajectories. The main idea is to replace a precise location by a rectangle enclosing a part of the trajectory. The size of the rectangles may vary. The authors propose a new method based on grids. The space is sliced in a set of 2D grids. Depending on the expected level of anonymity, the size of the used grid varies (the smaller the cell, the less privacy). They apply their solution to the problem of frequent route mining. They provide the experimental results of a real implementation.

Chapter 9: *Privacy and Anonymity in Location Data Management*

This chapter focuses on a similar problem than the previous chapter. Nevertheless, it focuses on a specific use: online privacy, i.e. the database is continuously evolving. Interestingly, this chapter presents a detailed threat model with a six parameters adversary model and briefly present asset of defense techniques. The literature summary is excellent and comprehensive. The last section details two emerging problems. How to preserve the anonymity of requester's identity when using LSB against historical attacks? How to preserve location anonymity for LSB that trigger the proximity of friends? This was one of the best chapters.

Chapter 10: *Privacy Preservation on Time Series*

This chapter focuses on time series and the perturbation techniques used to anonymize them. The use of techniques inherited from signal processing such as wavelets and digital filters may disclose sensitive information. For instance, white noise perturbations added to a signal can be easily removed through a simple compression. They introduce a new class of perturbation that adapts its perturbation to the original series. The idea is to make the perturbation in the transformed domain on the coefficients that are essential. This cross fertilization, using from a different domain, is extremely appealing nut requires some familiarity with signal processing.

Chapter 11: *A Segment-Based Approach to Preserve Privacy in Time Series Data Mining*

Like the previous chapter, the focus of this short chapter is on protecting time series. Nevertheless, the approach is different. Each time series is split into equal segments. Noise is added to each segment. The noise depends on the segment. This should allow defeating usual noise filtering.

Chapter 12: *A Survey of Challenges and Solutions for Privacy in Clinical Genomics Data Mining*

This chapter first presents the applicable regulations and standard practices (HiPAA, IRBs…). The key challenges are to avoid re-identification of de-identified genomics database. This is especially difficult since if the attacker has access to the genotypes and the list of diseases, she may guess as some diseases result from given DNA perturbations. The chapter presents an exhaustive description of attacks and countermeasures by preventing linkage between genomes and identities (using their own algorithm TRANON), preventing uniqueness of genomics via k-anonymity (using their own algorithm DNALA), or using secure multiparty computing via homomorphic encryption. If the reader is not familiar with the medical domain (as in my case), it is difficult to judge the value of the proposals.

Chapter 13: *Privacy-Aware Health Information Sharing*

Interestingly, this chapter presents a real use case. The Hong Kong Red Cross wants to perform data mining on the complete patient records with usage of blood. This database has a large dimension (over 60). The authors present countermeasures against record linkage and attribute linkage. They propose a framework for k-anonymity using quasi identifiers. They first tackle the problem for classification analysis. They apply the top down refinement algorithm to this framework to prevent record linkage using the real database. They benchmark the quality of some classifiers on the full database and their anonymized database. They briefly tackle the same framework for cluster analysis.

Chapter 14: *Issues with Privacy Preservation in Query Log Mining*

In 2006, in a famous paper, journalists from the New York Times demonstrated that it was possible to identify users from an AOL anonymized query log. This chapter explores the corresponding issue. First, the authors define what private data is using six criteria. They briefly explore the associated risks and corresponding usual countermeasures. It is an excellent introduction paper to the topic.

Chapter 15: *Preserving Privacy in Web Recommender Systems*

The chapter first introduces the topic and presents the two types of risks: data collection and breach privacy through recommendation. Then, it explores seven methods of recommendation methods. The last section presents an enhancement of a method the authors published previously. The new method assumes that the sensitive data will remain in the user's browser.

Chapter 16: *The Social Web and Privacy: Practices, Reciprocity and Conflict Detection in Social Networks*

This chapter mainly focuses on highlighting what privacy means for social networks. Privacy means protecting personal data and thwarting information leakage by exploiting relational information and transitive control. Indeed, these two characteristics of social networks may create inconsistency and conflicts which may be exploited to gain knowledge. The authors study conflict analysis in four use cases. The last section proposes to use data mining as a mirror of potential attackers to increase the users' awareness about privacy.

Chapter 17: *Privacy Protection of Personal Data in Social Networks*

This chapter tackles the issue of privacy in Social Networks from a radical different point of view: management of access control. First, it explains the challenges and highlights the difference between depth in the graph and trust level. Then, it carefully presents the semantic-based access control. The next section explores privacy challenges on two specific private data: the relationship itself between individuals and the associated trust level. Then, the chapter briefly explores the attacks by inference.

Chapter 18: *Analyzing Private Network Data*

This chapter analyses privacy from the point of view of network topology. A network, being social or not, is a graph. From this graph, it is possible to learn valuable information. The chapter presents a detailed overview of attacks that may reidentify nodes or disclose connections on anonymized networks. Next section describes how to protect through either by adding connections between nodes to generalize, clustering nodes to hide local topology, or adding random nodes and connections. The last section presents an alternative solution. Rather than publishing an anonymized network, the data owner answers queries that it may modify to enhance privacy.

Recommendation

This book suffers from a clear positioning. It is neither an introduction to privacy for data mining, nor a textbook on the subject. It is merely a collection of independent papers, at the exception of the section dedicated web usage and social networks (chapter 16 to 18) that seem to have some coordination. Thus, the book is a mix of papers which present state of the art of the field, and papers which present the research results of the authors. As usual with this type of book, there is a lot of redundancy between the papers. For instance, no paper takes advantage of the chapters of the first section that introduced concepts such as k-anonymity, perturbation...

As the goal of the book is not clear, it is difficult to state who should read it. If the reader is looking for an introduction to the domain of privacy and related techniques, she should look elsewhere. If the reader is looking for the latest advances on the topic, she should rather read proceedings of related conferences. The book demonstrates that the concept of privacy is highly context dependent. As the book explores different domains, the reader discovers many aspects of privacy. Albeit the presented solutions employ the same general concepts, they are all different in the implementation. The book demonstrates that anonymizing data is context dependent, extremely difficult, and still a domain in its infancy. If the reader is looking for a glimpse to this diversity, then this is a good book.

Conclusion when finishing reading the book, there is still a long way before we will able to ensure good privacy.

The reviewer is VP Security Systems and Technologies at Technicolor.