Chapter 1

An Introduction to Privacy-Preserving Data Mining

Charu C. Aggarwal
IBM T. J. Watson Research Center
Hawthorne, NY 10532
charu@us.ibm.com

Philip S. Yu
University of Illinois at Chicago
Chicago, IL 60607
psyu@cs.uic.edu

Abstract
The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy. While the topic of privacy has been traditionally studied in the context of cryptography and information-hiding, recent emphasis on data mining has lead to renewed interest in the field. In this chapter, we will introduce the topic of privacy-preserving data mining and provide an overview of the different topics covered in this book.

Keywords: Privacy-preserving data mining, privacy, randomization, k-anonymity.

1.1 Introduction
The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and \( k \)-anonymity [1, 4, 16] have been suggested in recent years in order to perform privacy-preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar. This book will try to explore different topics from the perspective of
different communities, and will try to give a fused idea of the work in different communities.

The key directions in the field of privacy-preserving data mining are as follows:

- **Privacy-Preserving Data Publishing:** These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization [1], \(k\)-anonymity [16, 7], and \(l\)-diversity [11]. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining [15]. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

- **Changing the results of Data Mining Applications to preserve privacy:** In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

- **Query Auditing:** Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries. Methods for perturbing the output of queries are discussed in [8], whereas techniques for restricting queries are discussed in [9, 13].

- **Cryptographic Methods for Distributed Privacy:** In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information. A survey of such methods may be found in [14].

- **Theoretical Challenges in High Dimensionality:** Real data sets are usually extremely high dimensional, and this makes the process of privacy-preservation extremely difficult both from a computational and effectiveness point of view. In [12], it has been shown that optimal \(k\)-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be