Chapter 7
Privacy Preserving Data Mining:
How Far Can We Go?

Aris Gkoulalas-Divanis
Vanderbilt University, USA

Vassilios S. Verykios
University of Thessaly, Greece

ABSTRACT

Since its inception in 2000, privacy preserving data mining has gained increasing popularity in the data mining research community. This line of research can be primarily attributed to the growing concern of individuals, organizations and the government regarding the violation of privacy in the mining of their data by the existing data mining technology. As a result, a whole new body of research was introduced to allow for the mining of data, while at the same time prohibiting the leakage of any private and sensitive information. In this chapter, the authors introduce the readers to the field of privacy preserving data mining; they discuss the reasons that led to its inception, the most prominent research directions, as well as some important methodologies per direction. Following that, the authors focus their attention on very recently investigated methodologies for the offering of privacy during the mining of user mobility data. In the end of the chapter, they provide a roadmap along with potential future research directions both with respect to the field of privacy-aware mobility data mining and to privacy preserving data mining at large.

INTRODUCTION

The significant advances in data collection and data storage technologies have provided the means for the inexpensive storage of enormous amounts of data in data warehouses that reside in companies and public organizations. Despite the benefit of using this data per se (e.g. for maintaining up to date profiles of the customers and record of their recent or historical purchases, maintaining an inventory of the available products, as well as their quantities and price, etc), the mining of these datasets with the existing data mining tools can reveal invaluable knowledge that was unknown to the data holder beforehand.

The extracted knowledge patterns can provide insight to the data holders and at the same time can be invaluable in tasks such as decision making and

DOI: 10.4018/978-1-60566-906-9.ch007

strategic planning. Moreover, private companies are often willing to collaborate with other entities who conduct similar business, towards the mutual benefit of their businesses. Significant knowledge patterns can be derived and shared among the collaborative partners with respect to the collective mining of their datasets. Furthermore, public sector organizations and civilian federal agencies usually have to share a portion of their collected data or knowledge with other organizations having a similar purpose, or even make this data and knowledge public. For example, the National Institute of Health (NIH) endorses research that leads to significant findings which improve human health and provides a set of guidelines which sanction the sharing of NIH-supported research findings with research institutions.

As it becomes evident, there exists an extended set of application scenarios in which information or knowledge derived from the data has to be shared with other (possibly untrusted) entities. Public agencies for example collect data for different purposes like population surveys, epidemiological and clinical studies, as well as various other social and economic experiments to answer a variety of problems that disturb the society as a whole. The sharing of data and/or knowledge may come at a cost to privacy, primarily due to two reasons: (a) if the data refers to individuals (e.g. as in customers' market basket data, medical data, preferences data and the like) then the disclosure of this data or any knowledge extracted from the data can potentially violate the privacy of the individuals if their identity is revealed to untrusted third parties, and (b) if the data regards business (or organizational) information, then the disclosure of this data or any knowledge extracted from the data may potentially reveal sensitive trade secrets, whose knowledge can provide a significant advantage to business competitors and thus can cause the data holder to lose business over his/her peers. The aforementioned privacy issues in the course of data mining are amplified due to the fact that untrusted entities (adversaries and data terrorists) may utilize other external and publicly available sources of information (e.g. the yellow pages, public reports) in conjunction with the released data or knowledge, in order to reveal even more protected sensitive information.

BACKGROUND

Since the pioneering work of Agrawal & Srikant (2000) and Lindell & Pinkas (2000), several approaches have been proposed for the offering of privacy in data mining. Most existing approaches can be classified along two broad categories: (a) methodologies that protect the sensitive data itself in the mining process, and (b) methodologies that protect the sensitive data mining results (i.e. the extracted knowledge patterns) that were produced by the application of data mining. The first category refers to methodologies that apply perturbation, sampling, generalization/suppression, transformation, etc. techniques to the original datasets in order to generate their sanitized counterparts that can be safely disclosed to untrusted third parties. The goal of this category of approaches is to enable the data miner to get accurate data mining results when is not provided with the real data.

As part of former category we highlight methodologies that have been proposed to enable a number of data holders to collectively mine their data without having to reveal their datasets to each other. On the other hand, the second category deals with distortion and blocking techniques that prohibit the disclosure of sensitive knowledge patterns derived through the application of data mining algorithms, as well as techniques for downgrading the effectiveness of the classifiers in classification tasks, such that they do not reveal any sensitive knowledge.

Vaidya, Clifton & Zhu (2006), and Aggarwal & Yu (2008) provide the different research directions that were investigated over the past eight years