IDEA GROUP PUBLISHING **ITJ3547**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

This paper appears in the publication, **International Journal of Information Security and Privacy**, Volume 1, Issue 1
edited by **Hamid R. Nemati © 2007, Idea Group Inc.**

Privacy-Preserving Data Mining and the Need for Confluence of Research and Practice

Lixin Fu, The University of North Carolina at Greensboro, USA

Hamid Nemati, The University of North Carolina at Greensboro, USA

Fereidoon Sadri, The University of North Carolina at Greensboro, USA

ABSTRACT

Privacy-preserving data mining (PPDM) refers to data mining techniques developed to protect sensitive data while allowing useful information to be discovered from the data. In this article, we review PPDM and present a broad survey of related issues, techniques, measures, applications, and regulation guidelines. We observe that the rapid pace of change in information technologies available to sustain PPDM has created a gap between theory and practice. We posit that without a clear understanding of the practice, this gap will be widening which, ultimately, will be detrimental to the field. We conclude by proposing a comprehensive research agenda intended to bridge the gap relevant to practice and as a reference basis for the future related legislation activities.

Keywords: *data mining; fair information practices; privacy laws; privacy preserving data mining*

INTRODUCTION

Technological advances, decreased costs of hardware and software, and the World Wide Web revolution have allowed for vast amounts of data to be generated, collected, stored, processed, analyzed, distributed, and used at an ever-increasing rate by organizations and governmental agencies. According to a survey by U.S. Department of Commerce, an increasing number of Americans are going online and engaging in several online activities, including online purchases and conducting banking online. The growth in Internet usage and e-commerce has offered businesses and governmental agencies the opportunity to collect and analyze information in ways never previously imagined. "Enormous amounts of consumer data have long been available through off-line sources such as credit card transactions, phone orders, warranty cards, applications, and a host of other traditional methods. What the digital revolution has done is increase the efficiency and effectiveness with which such information can be collected and put to use" (Adkinson, Eisenach, & Lenard, 2002).

Simultaneously, there is a growing awareness that by leveraging their data resources to develop and deploy data mining technologies

to enhance their decision-making capabilities, organizations can gain and sustain a competitive advantage (Eckerson & Watson, 2001). If correctly deployed, Data Mining (DM) offers organizations an indispensable decision-enhancing process that optimizes resource allocation and exploits new opportunities by transforming data into valuable knowledge (Nemati & Barko, 2001). Correctly deploying data mining has the potential of significantly increasing a company's profits and reducing its costs by helping to identify areas of potential business, or areas that the company needs to focus its attention on, or areas that should be discontinued because of poor sales or returns over a period of time. For example, data mining can identify customer buying patterns and preferences which would allow for a better management of inventory and new merchandising opportunities. However, when data contains personally identifiable attributes, and if data mining is used in the wrong context, it can be very harmful to individuals. Data mining may "pose a threat to privacy" in the sense that sensitive personal data may be exposed directly, or discovered patterns can reveal confidential personal attributes about individuals, or classify individuals into categories, revealing in that way confidential personal information with certain probability. Moreover, such patterns may lead to generation of stereotypes, raising very sensitive and controversial issues, especially if they involve attributes such as race, gender, or religion. An example is the debate about studies of "intelligence across different races" (Estivill-Castro, Brankovic, & Dowe, 1999). As another example, individual patient medical records are stored in electronic databases by government and private medical providers (Hodge, Gostin, & Jacobson, 1999). The proliferation of medical databases within the healthcare information infrastructure presents significant benefits for medical providers and patients, including enhanced patient autonomy, improved clinical treatment, advances in health research and public health surveillance (Hodge et al., 1999). However, use and mining of this type of data presents a significant risk of privacy. Therefore, not only protecting the confidentiality of personally-identifiable health data is critical, but also insufficient protections of what could be mined from it can subject the individuals to possible embarrassment, social stigma, and discrimination (Hodge et al., 1999).

The significance of data security and privacy has not been lost to the data mining research community, as was revealed in Nemati Barko (2001), of the major industry predictions that are expected to be key issues in the future. Chiefly among them are concerns over the security of what is collected and the privacy violations of what is discovered (Culnan, 1993; Margulis, 1977; Mason, 1986; Milberg, Smith, & Kallman, 1995, Smith, 1993; Smith, Milberg, & Burke, 1996). About 80 percent of survey respondents expect privacy implications of data mining of consumer data to be a significant issue (Nemati & Barko, 2001).

Recently, research in privacy as well as that of data mining has garnered considerable interest among academic and practitioners' communities from diverse perspectives, for example, technical, behavioral, sociological, governmental, and organizational perspectives. Although there is an extensive pool of literature that addresses many aspects of both privacy and data mining, it is often unclear as to how this literature relates and integrates to define an integrated privacy preserving data mining research discipline. Part of the problem is that there is no clear framework that defines the privacy-preserving data mining paradigm for research. As a result, privacy-preserving data mining research seems to be fluid and often fragmented. Hence, it is difficult to articulate what defines and bounds PPDM research and to identify what research streams exist within it. For academics, these issues can become problematic when seeking to define the contribution of intellectual activities in this discipline. Given the importance of this area of research, and the need for its viability as an increasingly-distinct discipline, a laudable move toward a tighter and conceptually better-developed research has recently become more evident. This trend has clearly been beneficial to the development