# Privacy-Preserving Collaborative Data Mining

Justin Zhan[1], LiWu Chang[2], and Stan Matwin[3]

[1] School of Information Technology & Engineering, University of Ottawa, Canada
zhizhan@site.uottawa.ca
[2] Center for High Assurance Computer Systems, Naval Research Laboratory, USA
lchang@itd.nrl.navy.mil
[3] School of Information Technology & Engineering, University of Ottawa, Canada
stan@site.uottawa.ca

**Summary.** Privacy-preserving data mining is an important issue in the areas of data mining and security. In this paper, we study how to conduct association rule mining, one of the core data mining techniques, on private data in the following scenario: Multiple parties, each having a private data set, want to jointly conduct association rule mining without disclosing their private data to other parties. Because of the interactive nature among parties, developing a secure framework to achieve such a computation is both challenging and desirable. In this paper, we present a secure framework for multiple parties to conduct privacy-preserving association rule mining.

**Key Words:**

privacy, security, association rule mining, secure multi-party computation.

## 1 INTRODUCTION

Business successes are no longer the result of an individual toiling in isolation; rather successes are dependent upon collaboration, team efforts, and partnership. In the modern business world, collaboration becomes especially important because of the mutual benefit it brings. Sometimes, such a collaboration even occurs among competitors, or among companies that have conflict of interests, but the collaborators are aware that the benefit brought by such a collaboration will give them an advantage over other competitors. For this kind of collaboration, data's privacy becomes extremely important: all the parties of the collaboration promise to provide their private data to the collaboration, but neither of them wants each other or any third party to learn much about their private data.

This paper studies a very specific collaboration that becomes more and more prevalent in the business world. The problem is the collaborative data

mining. Data mining is a technology that emerges as a means for identifying patterns and trends from a large quantity of data. The goal of our studies is to develop technologies to enable multiple parties to conduct data mining collaboratively without disclosing their private data.

In recent times, the explosion in the availability of various kinds of data has triggered tremendous opportunities for collaboration, in particular collaboration in data mining. The following is some realistic scenarios:

1. Multiple competing supermarkets, each having an extra large set of data records of its customers' buying behaviors, want to conduct data mining on their joint data set for mutual benefit. Since these companies are competitors in the market, they do not want to disclose too much about their customers' information to each other, but they know the results obtained from this collaboration could bring them an advantage over other competitors.
2. Several pharmaceutical companies, each have invested a significant amount of money conducting experiments related to human genes with the goal of discovering meaningful patterns among the genes. To reduce the cost, the companies decide to join force, but neither wants to disclose too much information about their raw data because they are only interested in this collaboration; by disclosing the raw data, a company essentially enables other parties to make discoveries that the company does not want to share with others.

To use the existing data mining algorithms, all parties need to send their data to a trusted central place (such as a super-computing center) to conduct the mining. However, in situations with privacy concerns, the parties may not trust anyone. We call this type of problem the *Privacy-preserving Collaborative Data Mining* (PCDM) problem. For each data mining problem, there is a corresponding PCDM problem. Fig.1 shows how a traditional data mining problem could be transformed to a PCDM problem (this paper only focuses on the heterogeneous collaboration (Fig.1.c))(heterogeneous collaboration means that each party has different sets of attributes. Homogeneous collaboration means that each party has the same sets of attributes.)

Generic solutions for any kind of secure collaborative computing exist in the literature [5]. These solutions are the results of the studies of the Secure Multi-party Computation problem [10, 5], which is a more general form of secure collaborative computing. However, none of the proposed generic solutions is practical; they are not scalable and cannot handle large-scale data sets because of the prohibitive extra cost in protecting data's privacy. Therefore, practical solutions need to be developed. This need underlies the rationale for our research.

Data mining includes a number of different tasks, such as association rule mining, classification, and clustering. This paper studies the association rule mining problem. The goal of association rule mining is to discover meaningful association rules among the attributes of a large quantity of data. For example,