Preface

Since its inception, privacy preserving data mining has been an active research area of increasing popularity in the data mining community. This line of research investigates the side-effects of the existing data mining technology that originate from the penetration into the privacy of individuals and organizations. From a general point of view, privacy issues related to the application of data mining can be classified into two main categories, namely *data hiding* and *knowledge hiding*. Data hiding methodologies are related to the data per se, aiming to remove confidential or private information from the data prior to its publication. Knowledge hiding methodologies, on the other hand, are concerned with the sanitization of data leading to the disclosure of confidential and private knowledge, when the data is mined by the existing data mining tools for knowledge patterns.

In this book, we provide an extensive survey on a specific class of privacy preserving data mining methods that belong to the knowledge hiding thread and are collectively known as *association rule hiding* methods. "Association rule hiding" (a term commonly used for brevity instead of the longer title "frequent itemset and association rule hiding") has been mentioned for the first time in 1999 in a workshop paper that was presented by Atallah et al. In this work, the authors tried to apply general ideas about the implications of data mining in security and privacy of information — first discussed by Clifton and Marks in 1996 — to the association rule mining framework proposed by Agrawal and Srikant. Clifton and Marks, following the suggestions of D.E. O'Leary (1991) — who was the very first to point out the security and privacy breaches that originate from data mining algorithms — indicated the need to consider different data mining approaches under the prism of preventing the privacy of information, and proposed different ways to accomplish this. Since then, a large body of research emerged that involved novel approaches for the hiding of sensitive association rules from within the data. Due to the combinatorial nature of the problem, the proposed methodologies span from simple, time and memory efficient heuristics and border-based approaches, to exact hiding algorithms that offer guarantees on the quality of the computed hiding solution at an increased, however, computational complexity cost. The focus of this book is mostly towards the latter type of approaches since it is also the most recent.

Book Organization

The book consists of 21 chapters, which are organized in four parts. Each part ends with a brief summary that overviews the covered material in the corresponding part. We tried to keep each chapter of the book self-contained so as to provide maximum reading flexibility.

The first part of the book presents some fundamental concepts for the understanding of the problem of association rule hiding as well as the key classes of proposed solution methodologies. It is composed of five chapters. Chapter 1 introduces the reader to association rule hiding and motivates this line of research. Next, Chapter 2 formally sets out the problem and provides the necessary notation and terminology for its proper understanding, while Chapter 3 sheds light on the different classes of association rule hiding methodologies that have been investigated over the years. Following that, Chapter 4 briefly discusses privacy preserving methodologies in related areas of research and specifically in the areas of classification, clustering and sequence mining. Last, Chapter 5 summarizes the contents of the first part.

The second part of the book contains three chapters and covers the heuristic class of association rule hiding methodologies. Two main directions of heuristic methodologies have been investigated over the years: support-based and confidence-based distortion schemes (presented in Chapter 6) that operate by including or excluding specific items in/from transactions of the original database, and support-based and confidence-based blocking approaches (covered in Chapter 7) that replace original values with question marks, reducing in this way the confidence of attackers regarding the existence (or nonexistence) of specific items in transactions of the original database. Chapter 8 summarizes the contents of the second part of the book.

The use of the theory of borders of the frequent itemsets to support association rule hiding is the key principle behind the border-based class of approaches, presented in detail in the third part of the book. This part comprises four chapters. Chapter 9 elucidates the process of border revision and presents a set of algorithms that can be directly applied to the association rule mining framework to allow for the efficient computation of the original and the revised borders of the frequent itemsets. Following that, Chapters 10 and 11 present two specific border-based methodologies for association rule hiding and demonstrate their way of operation. A short summary of this part is given in Chapter 12.

The last part of the book is devoted to exact association rule hiding methodologies, which are the only ones to offer guarantees on the quality of the identified hiding solution. Since this line of research is also the most recent one, the detail of presentation is intentionally finer in favor of this class of approaches. This part is comprised of seven chapters. Chapter 13 presents the first work to elevate from pure heuristics to exact knowledge hiding by somewhat combining the two worlds. Next, Chapters 14, 15 and 16 study in detail three exact association rule hiding methodologies that formulate the problem of association rule hiding as a pure optimization problem and offer quality guarantees on the identified hiding solution. A notable drawback of exact hiding methodologies, mainly attributed to their optimization nature, is their high computational and memory requirements. Chapter 17 discusses

a decomposition and parallelization framework that has been recently developed to ameliorate these disadvantages by effectively decomposing large optimization problems into smaller subproblems that can be solved concurrently, without however sacrificing the quality of the computed hiding solution. Following that, Chapter 18 presents a systematic layered approach for the quantification of privacy that is offered by the exact hiding algorithms. The proposed approach allows data owners to decide on the level of privacy they wish as a tradeoff of the distortion that is induced to the database by the hiding process. In this way, the exact algorithms can effectively shield all the sensitive association rules with the least possible damage to data utility. Chapter 19 summarizes the contributions of this part.

The Epilogue comprises two chapters that are wrapping up the discussion presented in this book. In Chapter 20, we give a summary of the presented material and emphasize on the most important topics that were covered. Finally, Chapter 21 elaborates on a number of interesting and promising directions for future research in the area of association rule hiding.

Intended Audience

We believe that this book will be suitable to course instructors, undergraduate and postgraduate students studying association rule hiding, knowledge hiding and privacy preserving data mining at large. Moreover, it is expected to be a valuable companion to researchers and professionals working in this research area as it provides an overview of the current research accomplishments by presenting them under a new perspective, building its way up from theory to practice and from simple heuristic methodologies to more advanced exact knowledge hiding approaches that have been proposed for association rule hiding. Finally, practitioners working on the development of association rule hiding methodologies for database systems can benefit from this book by using it as a reference guide to decide upon the hiding approach that best fits their needs as well as gain insight on the eccentricities and peculiarities of the existing methodologies.

How to Study This Book

The order of presentation is also the proposed reading order of the material. This way, the reader can start from the basics and incrementally build his or her way up to more advanced topics, including the different classes of hiding approaches, the underlying principles of the proposed methodologies and their main differences. However, based on the reader's expertise in the area, it is possible to focus directly on the topic of interest and thus skip either the first part of the book or parts devoted to other classes of approaches. Evidently, if the reader wishes to go deep into the details of a certain subject topic, then the provided references should be consulted.

More precisely, undergraduate students can focus on the first two parts of the book, effectively gain a good understanding of the fundamental concepts related to association rule hiding and grasp the main characteristics of some of the most popular heuristic methodologies in this area. Postgraduate students and researchers will find the third and fourth parts of the book to be more interesting as they cover more recent developments in association rule hiding. Specifically the fourth part of the book reviews the most recent line of research in association rule hiding, which is expected to radically benefit from the advances in the area of optimization techniques. Course instructors and researchers are encouraged to study all the material in order to select the most suitable parts of the book required for class presentation or further research in the area.

Acknowledgments

This work aims to summarize the most interesting research accomplishments that took place in the area of association rule hiding during the ten years of its existence. Since this is the first book in the market that is specifically targeted on association rule hiding, we would like to express our deep gratitude to Ahmed K. Elmagarmid, Chris Clifton, Yucel Saygin, Osmar R. Zaïane, Charu Aggarwal and Francesco Bonchi for cordially embracing our book proposal and for offering constructive comments that helped us improve the overall quality of the manuscript.

We are also indebted to Susan Lagerstrom-Fife and Jennifer Maurer from Springer, for their great support towards the preparation and completion of this work. Their editing suggestions were valuable to improving the organization, readability and appearance of the manuscript.

Finally, we would like to express our deep love and gratitude to our families for their understanding throughout the duration of this project.

We really hope that this book will serve as a valuable resource to researchers, graduate students and professors interested in the area of association rule hiding and privacy preserving data mining at large, as well as a motivating companion to senior undergraduate students who wish to study the theory and methods in this research area.

Aris Gkoulalas-Divanis, Vanderbilt University, Nashville, USA.
Vassilios S. Verykios, University of Thessaly, Volos, GREECE.

January 2010