

Privacy-Preserving Data Mining

Models and Algorithms

ADVANCES IN DATABASE SYSTEMS

Volume 34

Series Editors

Ahmed K. Elmagarmid

Purdue University
West Lafayette, IN 47907

Amit P. Sheth

Wright State University
Dayton, Ohio 45435

Other books in the Series:

SEQUENCE DATA MINING, Guozhu Dong, Jian Pei; ISBN: 978-0-387-69936-3

DATA STREAMS: Models and Algorithms, edited by Charu C. Aggarwal; ISBN: 978-0-387-28759-1

SIMILARITY SEARCH: The Metric Space Approach, P. Zezula, G. Amato, V. Dohnal, M. Batko;
ISBN: 0-387-29146-6

STREAM DATA MANAGEMENT, Nauman Chaudhry, Kevin Shaw, Mahdi Abdelguerfi; ISBN:
0-387-24393-3

FUZZY DATABASE MODELING WITH XML, Zongmin Ma; ISBN: 0-387-24248-1

MINING SEQUENTIAL PATTERNS FROM LARGE DATA SETS, Wei Wang and Jiong Yang;
ISBN: 0-387-24246-5

ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS, Yan-
nis Manolopoulos, Alexandros Nanopoulos, Eleni Tousidou; ISBN: 1-4020-7425-5

ADVANCES IN DIGITAL GOVERNMENT: Technology, Human Factors, and Policy, edited by
William J. McIver, Jr. and Ahmed K. Elmagarmid; ISBN: 1-4020-7067-5

INFORMATION AND DATABASE QUALITY, Mario Piattini, Coral Calero and Marcela Genero;
ISBN: 0-7923-7599-8

DATA QUALITY, Richard Y. Wang, Mostapha Ziad, Yang W. Lee; ISBN: 0-7923-7215-8

THE FRACTAL STRUCTURE OF DATA REFERENCE: Applications to the Memory Hierarchy,
Bruce McNutt; ISBN: 0-7923-7945-4

SEMANTIC MODELS FOR MULTIMEDIA DATABASE SEARCHING AND BROWSING, Shu-
Ching Chen, R.L. Kashyap, and Arif Ghaffoor; ISBN: 0-7923-7888-1

**INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA: A Metadata-
based Approach**, Vipul Kashyap, Amit Sheth; ISBN: 0-7923-7883-0

DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS, Kian-Lee Tan and
Beng Chin Ooi; ISBN: 0-7923-7866-0

MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure,
Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic; ISBN: 0-7923-7840-7

ADVANCED DATABASE INDEXING, Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras;
ISBN: 0-7923-7716-8

MULTILEVEL SECURE TRANSACTION PROCESSING, Vijay Athuri, Sushil Jajodia, Binto
George; ISBN: 0-7923-7702-8

FUZZY LOGIC IN DATA MODELING, Guoqing Chen; ISBN: 0-7923-8253-6

PRIVACY-PRESERVING DATA MINING: Models and Algorithms, edited by Charu C. Aggarwal
and Philip S. Yu; ISBN: 0-387-70991-8

Privacy-Preserving Data Mining

Models and Algorithms

Edited by

Charu C. Aggarwal

IBM T.J. Watson Research Center, USA

and

Philip S. Yu

University of Illinois at Chicago, USA

 Springer

Editors:

Charu C. Aggarwal
IBM Thomas J. Watson Research Center
19 Skyline Drive
Hawthorne NY 10532
charu@us.ibm.com

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
854 South Morgan Street
Chicago, IL 60607-7053
psyu@cs.uic.edu

Series Editors

Ahmed K. Elmagarmid
Purdue University
West Lafayette, IN 47907

Amit P. Sheth
Wright State University
Dayton, Ohio 45435

ISBN 978-0-387-70991-8

e-ISBN 978-0-387-70992-5

DOI 10.1007/978-0-387-70992-5

Library of Congress Control Number: 2007943463

© 2008 Springer Science+Business Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

In recent years, advances in hardware technology have led to an increase in the capability to store and record personal data about consumers and individuals. This has led to concerns that the personal data may be misused for a variety of purposes. In order to alleviate these concerns, a number of techniques have recently been proposed in order to perform the data mining tasks in a privacy-preserving way. These techniques for performing privacy-preserving data mining are drawn from a wide array of related topics such as data mining, cryptography and information hiding. The material in this book is designed to be drawn from the different topics so as to provide a good overview of the important topics in the field.

While a large number of research papers are now available in this field, many of the topics have been studied by different communities with different styles. At this stage, it becomes important to organize the topics in such a way that the relative importance of different research areas is recognized. Furthermore, the field of privacy-preserving data mining has been explored independently by the cryptography, database and statistical disclosure control communities. In some cases, the parallel lines of work are quite similar, but the communities are not sufficiently integrated for the provision of a broader perspective. This book will contain chapters from researchers of all three communities and will therefore try to provide a balanced perspective of the work done in this field.

This book will be structured as an edited book from prominent researchers in the field. Each chapter will contain a survey which contains the key research content on the topic, and the future directions of research in the field. Emphasis will be placed on making each chapter self-sufficient. While the chapters will be written by different researchers, the topics and content is organized in such a way so as to present the most important models, algorithms, and applications in the privacy field in a structured and concise way. In addition, attention is paid in drawing chapters from researchers working in different areas in order to provide different points of view. Given the lack of structurally organized information on the topic of privacy, the book will provide insights which are not easily accessible otherwise. A few chapters in the book are not surveys, since the corresponding topics fall in the emerging category, and enough material is

not available to create a survey. In such cases, the individual results have been included to give a flavor of the emerging research in the field. It is expected that the book will be a great help to researchers and graduate students interested in the topic. While the privacy field clearly falls in the emerging category because of its recency, it is now beginning to reach a maturation and popularity point, where the development of an overview book on the topic becomes both possible and necessary. It is hoped that this book will provide a reference to students, researchers and practitioners in both introducing the topic of privacy-preserving data mining and understanding the practical and algorithmic aspects of the area.

Contents

Preface	v
List of Figures	xvii
List of Tables	xxi
1	
An Introduction to Privacy-Preserving Data Mining	1
<i>Charu C. Aggarwal, Philip S. Yu</i>	
1.1. Introduction	1
1.2. Privacy-Preserving Data Mining Algorithms	3
1.3. Conclusions and Summary	7
References	8
2	
A General Survey of Privacy-Preserving Data Mining Models and Algorithms	11
<i>Charu C. Aggarwal, Philip S. Yu</i>	
2.1. Introduction	11
2.2. The Randomization Method	13
2.2.1 Privacy Quantification	15
2.2.2 Adversarial Attacks on Randomization	18
2.2.3 Randomization Methods for Data Streams	18
2.2.4 Multiplicative Perturbations	19
2.2.5 Data Swapping	19
2.3. Group Based Anonymization	20
2.3.1 The k -Anonymity Framework	20
2.3.2 Personalized Privacy-Preservation	24
2.3.3 Utility Based Privacy Preservation	24
2.3.4 Sequential Releases	25
2.3.5 The l -diversity Method	26
2.3.6 The t -closeness Model	27
2.3.7 Models for Text, Binary and String Data	27
2.4. Distributed Privacy-Preserving Data Mining	28
2.4.1 Distributed Algorithms over Horizontally Partitioned Data Sets	30
2.4.2 Distributed Algorithms over Vertically Partitioned Data	31
2.4.3 Distributed Algorithms for k -Anonymity	32

2.5.	Privacy-Preservation of Application Results	32
2.5.1	Association Rule Hiding	33
2.5.2	Downgrading Classifier Effectiveness	34
2.5.3	Query Auditing and Inference Control	34
2.6.	Limitations of Privacy: The Curse of Dimensionality	37
2.7.	Applications of Privacy-Preserving Data Mining	38
2.7.1	Medical Databases: The Scrub and Datafly Systems	39
2.7.2	Bioterrorism Applications	40
2.7.3	Homeland Security Applications	40
2.7.4	Genomic Privacy	42
2.8.	Summary	43
	References	43
3		
A	Survey of Inference Control Methods for Privacy-Preserving Data Mining	53
	<i>Josep Domingo-Ferrer</i>	
3.1.	Introduction	54
3.2.	A classification of Microdata Protection Methods	55
3.3.	Perturbative Masking Methods	58
3.3.1	Additive Noise	58
3.3.2	Microaggregation	59
3.3.3	Data Wapping and Rank Swapping	61
3.3.4	Rounding	62
3.3.5	Resampling	62
3.3.6	PRAM	62
3.3.7	MASSC	63
3.4.	Non-perturbative Masking Methods	63
3.4.1	Sampling	64
3.4.2	Global Recoding	64
3.4.3	Top and Bottom Coding	65
3.4.4	Local Suppression	65
3.5.	Synthetic Microdata Generation	65
3.5.1	Synthetic Data by Multiple Imputation	65
3.5.2	Synthetic Data by Bootstrap	66
3.5.3	Synthetic Data by Latin Hypercube Sampling	66
3.5.4	Partially Synthetic Data by Cholesky Decomposition	67
3.5.5	Other Partially Synthetic and Hybrid Microdata Approaches	67
3.5.6	Pros and Cons of Synthetic Microdata	68
3.6.	Trading off Information Loss and Disclosure Risk	69
3.6.1	Score Construction	69
3.6.2	R-U Maps	71
3.6.3	k -anonymity	71
3.7.	Conclusions and Research Directions	72
	References	73

<i>Contents</i>	ix
4	
Measures of Anonymity	81
<i>Suresh Venkatasubramanian</i>	
4.1. Introduction	81
4.1.1 What is Privacy?	82
4.1.2 Data Anonymization Methods	83
4.1.3 A Classification of Methods	84
4.2. Statistical Measures of Anonymity	85
4.2.1 Query Restriction	85
4.2.2 Anonymity via Variance	85
4.2.3 Anonymity via Multiplicity	86
4.3. Probabilistic Measures of Anonymity	87
4.3.1 Measures Based on Random Perturbation	87
4.3.2 Measures Based on Generalization	90
4.3.3 Utility vs Privacy	94
4.4. Computational Measures of Anonymity	94
4.4.1 Anonymity via Isolation	97
4.5. Conclusions and New Directions	97
4.5.1 New Directions	98
References	99
5	
<i>k</i> -Anonymous Data Mining: A Survey	105
<i>V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati</i>	
5.1. Introduction	105
5.2. <i>k</i> -Anonymity	107
5.3. Algorithms for Enforcing <i>k</i> -Anonymity	110
5.4. <i>k</i> -Anonymity Threats from Data Mining	117
5.4.1 Association Rules	118
5.4.2 Classification Mining	118
5.5. <i>k</i> -Anonymity in Data Mining	120
5.6. Anonymize-and-Mine	123
5.7. Mine-and-Anonymize	126
5.7.1 Enforcing <i>k</i> -Anonymity on Association Rules	126
5.7.2 Enforcing <i>k</i> -Anonymity on Decision Trees	130
5.8. Conclusions	133
Acknowledgments	133
References	134
6	
A Survey of Randomization Methods for Privacy-Preserving Data Mining	137
<i>Charu C. Aggarwal, Philip S. Yu</i>	
6.1. Introduction	137
6.2. Reconstruction Methods for Randomization	139
6.2.1 The Bayes Reconstruction Method	139
6.2.2 The EM Reconstruction Method	141
6.2.3 Utility and Optimality of Randomization Models	143

6.3.	Applications of Randomization	144
6.3.1	Privacy-Preserving Classification with Randomization	144
6.3.2	Privacy-Preserving OLAP	145
6.3.3	Collaborative Filtering	145
6.4.	The Privacy-Information Loss Tradeoff	146
6.5.	Vulnerabilities of the Randomization Method	149
6.6.	Randomization of Time Series Data Streams	151
6.7.	Multiplicative Noise for Randomization	152
6.7.1	Vulnerabilities of Multiplicative Randomization	153
6.7.2	Sketch Based Randomization	153
6.8.	Conclusions and Summary	154
	References	154
7		
	A Survey of Multiplicative Perturbation for Privacy-Preserving Data Mining	157
	<i>Keke Chen and Ling Liu</i>	
7.1.	Introduction	158
7.1.1	Data Privacy vs. Data Utility	159
7.1.2	Outline	160
7.2.	Definition of Multiplicative Perturbation	161
7.2.1	Notations	161
7.2.2	Rotation Perturbation	161
7.2.3	Projection Perturbation	162
7.2.4	Sketch-based Approach	164
7.2.5	Geometric Perturbation	164
7.3.	Transformation Invariant Data Mining Models	165
7.3.1	Definition of Transformation Invariant Models	166
7.3.2	Transformation-Invariant Classification Models	166
7.3.3	Transformation-Invariant Clustering Models	167
7.4.	Privacy Evaluation for Multiplicative Perturbation	168
7.4.1	A Conceptual Multidimensional Privacy Evaluation Model	168
7.4.2	Variance of Difference as Column Privacy Metric	169
7.4.3	Incorporating Attack Evaluation	170
7.4.4	Other Metrics	171
7.5.	Attack Resilient Multiplicative Perturbations	171
7.5.1	Naive Estimation to Rotation Perturbation	171
7.5.2	ICA-Based Attacks	173
7.5.3	Distance-Inference Attacks	174
7.5.4	Attacks with More Prior Knowledge	176
7.5.5	Finding Attack-Resilient Perturbations	177
7.6.	Conclusion	177
	Acknowledgment	178
	References	179
8		
	A Survey of Quantification of Privacy Preserving Data Mining Algorithms	183
	<i>Elisa Bertino, Dan Lin and Wei Jiang</i>	
8.1.	Introduction	184
8.2.	Metrics for Quantifying Privacy Level	186
8.2.1	Data Privacy	186

8.2.2	Result Privacy	191
8.3.	Metrics for Quantifying Hiding Failure	192
8.4.	Metrics for Quantifying Data Quality	193
8.4.1	Quality of the Data Resulting from the PPDM Process	193
8.4.2	Quality of the Data Mining Results	198
8.5.	Complexity Metrics	200
8.6.	How to Select a Proper Metric	201
8.7.	Conclusion and Research Directions	202
	References	202
9		
A Survey of Utility-based Privacy-Preserving Data Transformation Methods		207
<i>Ming Hua and Jian Pei</i>		
9.1.	Introduction	208
9.1.1	What is Utility-based Privacy Preservation?	209
9.2.	Types of Utility-based Privacy Preservation Methods	210
9.2.1	Privacy Models	210
9.2.2	Utility Measures	212
9.2.3	Summary of the Utility-Based Privacy Preserving Methods	214
9.3.	Utility-Based Anonymization Using Local Recoding	214
9.3.1	Global Recoding and Local Recoding	215
9.3.2	Utility Measure	216
9.3.3	Anonymization Methods	217
9.3.4	Summary and Discussion	219
9.4.	The Utility-based Privacy Preserving Methods in Classification Problems	219
9.4.1	The Top-Down Specialization Method	220
9.4.2	The Progressive Disclosure Algorithm	224
9.4.3	Summary and Discussion	228
9.5.	Anonymized Marginal: Injecting Utility into Anonymized Data Sets	228
9.5.1	Anonymized Marginal	229
9.5.2	Utility Measure	230
9.5.3	Injecting Utility Using Anonymized Marginals	231
9.5.4	Summary and Discussion	233
9.6.	Summary	234
	Acknowledgments	234
	References	234
10		
Mining Association Rules under Privacy Constraints		239
<i>Jayant R. Haritsa</i>		
10.1.	Introduction	239
10.2.	Problem Framework	240
10.2.1	Database Model	240
10.2.2	Mining Objective	241
10.2.3	Privacy Mechanisms	241
10.2.4	Privacy Metric	243
10.2.5	Accuracy Metric	245

10.3.	Evolution of the Literature	246
10.4.	The FRAPP Framework	251
10.4.1	Reconstruction Model	252
10.4.2	Estimation Error	253
10.4.3	Randomizing the Perturbation Matrix	256
10.4.4	Efficient Perturbation	256
10.4.5	Integration with Association Rule Mining	258
10.5.	Sample Results	259
10.6.	Closing Remarks	263
	Acknowledgments	263
	References	263
11		
	A Survey of Association Rule Hiding Methods for Privacy	267
	<i>Vassilios S. Verykios and Aris Gkoulalas-Divanis</i>	
11.1.	Introduction	267
11.2.	Terminology and Preliminaries	269
11.3.	Taxonomy of Association Rule Hiding Algorithms	270
11.4.	Classes of Association Rule Algorithms	271
11.4.1	Heuristic Approaches	272
11.4.2	Border-based Approaches	277
11.4.3	Exact Approaches	278
11.5.	Other Hiding Approaches	279
11.6.	Metrics and Performance Analysis	281
11.7.	Discussion and Future Trends	284
11.8.	Conclusions	285
	References	286
12		
	A Survey of Statistical Approaches to Preserving Confidentiality of Contingency Table Entries	291
	<i>Stephen E. Fienberg and Aleksandra B. Slavkovic</i>	
12.1.	Introduction	291
12.2.	The Statistical Approach Privacy Protection	292
12.3.	Datamining Algorithms, Association Rules, and Disclosure Limitation	294
12.4.	Estimation and Disclosure Limitation for Multi-way Contingency Tables	295
12.5.	Two Illustrative Examples	301
12.5.1	Example 1: Data from a Randomized Clinical Trial	301
12.5.2	Example 2: Data from the 1993 U.S. Current Population Survey	305
12.6.	Conclusions	308
	Acknowledgments	309
	References	309
13		
	A Survey of Privacy-Preserving Methods Across Horizontally Partitioned Data	313
	<i>Murat Kantarcioglu</i>	
13.1.	Introduction	313

13.2.	Basic Cryptographic Techniques for Privacy-Preserving Distributed Data Mining	315
13.3.	Common Secure Sub-protocols Used in Privacy-Preserving Distributed Data Mining	318
13.4.	Privacy-preserving Distributed Data Mining on Horizontally Partitioned Data	323
13.5.	Comparison to Vertically Partitioned Data Model	326
13.6.	Extension to Malicious Parties	327
13.7.	Limitations of the Cryptographic Techniques Used in Privacy-Preserving Distributed Data Mining	329
13.8.	Privacy Issues Related to Data Mining Results	330
13.9.	Conclusion	332
	References	332
14		
A	Survey of Privacy-Preserving Methods Across Vertically Partitioned Data	337
<i>Jaideep Vaidya</i>		
14.1.	Introduction	337
14.2.	Classification	341
	14.2.1 Naïve Bayes Classification	342
	14.2.2 Bayesian Network Structure Learning	343
	14.2.3 Decision Tree Classification	344
14.3.	Clustering	346
14.4.	Association Rule Mining	347
14.5.	Outlier detection	349
	14.5.1 Algorithm	351
	14.5.2 Security Analysis	352
	14.5.3 Computation and Communication Analysis	354
14.6.	Challenges and Research Directions	355
	References	356
15		
A	Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods	359
<i>Kun Liu, Chris Giannella, and Hillol Kargupta</i>		
15.1.	Introduction	360
15.2.	Definitions and Notation	360
15.3.	Attacking Additive Data Perturbation	361
	15.3.1 Eigen-Analysis and PCA Preliminaries	362
	15.3.2 Spectral Filtering	363
	15.3.3 SVD Filtering	364
	15.3.4 PCA Filtering	365
	15.3.5 MAP Estimation Attack	366
	15.3.6 Distribution Analysis Attack	367
	15.3.7 Summary	367
15.4.	Attacking Matrix Multiplicative Data Perturbation	369
	15.4.1 Known I/O Attacks	370
	15.4.2 Known Sample Attack	373
	15.4.3 Other Attacks Based on ICA	374

15.4.4	Summary	375
15.5.	Attacking k -Anonymization	376
15.6.	Conclusion	376
	Acknowledgments	377
	References	377
16		
	Private Data Analysis via Output Perturbation	383
	<i>Kobbi Nissim</i>	
16.1.	Introduction	383
16.2.	The Abstract Model – Statistical Databases, Queries, and Sanitizers	385
16.3.	Privacy	388
16.3.1	Interpreting the Privacy Definition	390
16.4.	The Basic Technique: Calibrating Noise to Sensitivity	394
16.4.1	Applications: Functions with Low Global Sensitivity	396
16.5.	Constructing Sanitizers for Complex Functionalities	400
16.5.1	k -Means Clustering	401
16.5.2	SVD and PCA	403
16.5.3	Learning in the Statistical Queries Model	404
16.6.	Beyond the Basics	405
16.6.1	Instance Based Noise and Smooth Sensitivity	406
16.6.2	The Sample-Aggregate Framework	408
16.6.3	A General Sanitization Mechanism	409
16.7.	Related Work and Bibliographic Notes	409
	Acknowledgments	411
	References	411
17		
	A Survey of Query Auditing Techniques for Data Privacy	415
	<i>Shubha U. Narar, Krishnaram Kenthapadi, Nina Mishra and Rajeev Motwani</i>	
17.1.	Introduction	415
17.2.	Auditing Aggregate Queries	416
17.2.1	Offline Auditing	417
17.2.2	Online Auditing	418
17.3.	Auditing Select-Project-Join Queries	426
17.4.	Challenges in Auditing	427
17.5.	Reading	429
	References	430
18		
	Privacy and the Dimensionality Curse	433
	<i>Charu C. Aggarwal</i>	
18.1.	Introduction	433
18.2.	The Dimensionality Curse and the k -anonymity Method	435
18.3.	The Dimensionality Curse and Condensation	441
18.4.	The Dimensionality Curse and the Randomization Method	446
18.4.1	Effects of Public Information	446
18.4.2	Effects of High Dimensionality	450
18.4.3	Gaussian Perturbing Distribution	450
18.4.4	Uniform Perturbing Distribution	455

18.5. The Dimensionality Curse and l -diversity	458
18.6. Conclusions and Research Directions	459
References	460
19	
Personalized Privacy Preservation	461
<i>Yufei Tao and Xiaokui Xiao</i>	
19.1. Introduction	461
19.2. Formalization of Personalized Anonymity	463
19.2.1 Personal Privacy Requirements	464
19.2.2 Generalization	465
19.3. Combinatorial Process of Privacy Attack	467
19.3.1 Primary Case	468
19.3.2 Non-primary Case	469
19.4. Theoretical Foundation	470
19.4.1 Notations and Basic Properties	471
19.4.2 Derivation of the Breach Probability	472
19.5. Generalization Algorithm	473
19.5.1 The Greedy Framework	474
19.5.2 Optimal SA-generalization	476
19.6. Alternative Forms of Personalized Privacy Preservation	478
19.6.1 Extension of k -anonymity	479
19.6.2 Personalization in Location Privacy Protection	480
19.7. Summary and Future Work	482
References	485
20	
Privacy-Preserving Data Stream Classification	487
<i>Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Rong She, and Jian Pei</i>	
20.1. Introduction	487
20.1.1 Motivating Example	488
20.1.2 Contributions and Paper Outline	490
20.2. Related Works	491
20.3. Problem Statement	493
20.3.1 Secure Join Stream Classification	493
20.3.2 Naive Bayesian Classifiers	494
20.4. Our Approach	495
20.4.1 Initialization	495
20.4.2 Bottom-Up Propagation	496
20.4.3 Top-Down Propagation	497
20.4.4 Using NBC	499
20.4.5 Algorithm Analysis	500
20.5. Empirical Studies	501
20.5.1 Real-life Datasets	502
20.5.2 Synthetic Datasets	504
20.5.3 Discussion	506
20.6. Conclusions	507
References	508
Index	511

List of Figures

5.1	Simplified representation of a private table	108
5.2	An example of domain and value generalization hierarchies	109
5.3	Classification of k -anonymity techniques [11]	110
5.4	Generalization hierarchy for $QI=\{\text{Marital_status}, \text{Sex}\}$	111
5.5	Index assignment to attributes <code>Marital_status</code> and <code>Sex</code>	112
5.6	An example of set enumeration tree over set $\mathcal{I} = \{1, 2, 3\}$ of indexes	113
5.7	Sub-hierarchies computed by Incognito for the table in Figure 5.1	114
5.8	Spatial representation (a) and possible partitioning (b)-(d) of the table in Figure 5.1	116
5.9	An example of decision tree	119
5.10	Different approaches for combining k -anonymity and data mining	120
5.11	An example of top-down anonymization for the private table in Figure 5.1	124
5.12	Frequent itemsets extracted from the table in Figure 5.1	127
5.13	An example of binary table	128
5.14	Itemsets extracted from the table in Figure 5.13(b)	128
5.15	Itemsets with support at least equal to 40 (a) and corresponding anonymized itemsets (b)	129
5.16	3-anonymous version of the tree of Figure 5.9	131
5.17	Suppression of occurrences in non-leaf nodes in the tree in Figure 5.9	132
5.18	Table inferred from the decision tree in Figure 5.17	132
5.19	11-anonymous version of the tree in Figure 5.17	132
5.20	Table inferred from the decision tree in Figure 5.19	133
6.1	Illustration of the Information Loss Metric	149
7.1	Using known points and distance relationship to infer the rotation matrix	175

9.1	A taxonomy tree on categorical attribute <i>Education</i>	221
9.2	A taxonomy tree on continuous attribute <i>Age</i>	221
9.3	Interactive graph	232
9.4	A decomposition	232
10.1	CENSUS ($\gamma = 19$)	261
10.2	Perturbation Matrix Condition Numbers ($\gamma = 19$)	262
13.1	Relationship between Secure Sub-protocols and Privacy Preserving Distributed Data Mining on Horizontally Partitioned Data	323
14.1	Two dimensional problem that cannot be decomposed into two one-dimensional problems	340
15.1	Wigner's semi-circle law: a histogram of the eigenvalues of $\frac{A+A'}{2\sqrt{2p}}$ for a large, randomly generated A	363
17.1	Skeleton of a simulatable private randomized auditor	423
18.1	Some Examples of Generalization for 2-Anonymity	435
18.2	Upper Bound of 2-anonymity Probability in an Non-Empty Grid Cell	439
18.3	Fraction of Data Points Preserving 2-Anonymity with Data Dimensionality (Gaussian Clusters)	440
18.4	Minimum Information Loss for 2-Anonymity (Gaussian Clusters)	445
18.5	Randomization Level with Increasing Dimensionality, Perturbation level = $8 \cdot \sigma^o$ (<i>UniDis</i>)	457
19.1	Microdata and generalization	462
19.2	The taxonomy of attribute <i>Disease</i>	463
19.3	A possible result of our generalization scheme	466
19.4	The voter registration list	468
19.5	Algorithm for computing personalized generalization	474
19.6	Algorithm for finding the optimal SA-generalization	478
19.7	Personalized k -anonymous generalization	480
20.1	Related streams / tables	489
20.2	The join stream	489
20.3	Example with 3 streams at initialization	496
20.4	After bottom-up propagations	498
20.5	After top-down propagations	499
20.6	UK road accident data (2001)	502
20.7	Classifier accuracy	503
20.8	Time per input tuple	503

20.9	Classifier accuracy vs. window size	505
20.10	Classifier accuracy vs. concept drifting interval	505
20.11	Time per input tuple vs. window size	506
20.12	Time per input tuple vs. blow-up ratio	506
20.13	Time per input tuple vs. number of streams	507

List of Tables

3.1	Perturbative methods vs data types. “X” denotes applicable and “(X)” denotes applicable with some adaptation	58
3.2	Example of rank swapping. Left, original file; right, rankswapped file	62
3.3	Non-perturbative methods vs data types	64
9.1a	The original table	209
9.2b	A 2-anonymized table with better utility	209
9.3c	A 2-anonymized table with poorer utility	209
9.4	Summary of utility-based privacy preserving methods	214
9.5a	3-anonymous table by global recoding	215
9.6b	3-anonymous table by local recoding	215
9.7a	The original table	223
9.8b	The anonymized table	223
9.9a	The original table	225
9.10b	The suppressed table	225
9.11a	The original table	229
9.12b	The anonymized table	229
9.13a	<i>Age</i> Marginal	229
9.14b	<i>(Education, AnnualIncome)</i> Marginal	229
10.1	CENSUS Dataset	260
10.2	Frequent Itemsets for $sup_{min} = 0.02$	260
12.1	Results of clinical trial for the effectiveness of an analgesic drug	302
12.2	Second panel has LP relaxation bounds, and third panel has sharp IP bounds for cell entries in Table 1.1 given $[R CST]$ conditional probability values	303
12.3	Sharp upper and lower bounds for cell entries in Table 12.1 given the $[CSR]$ margin, and LP relaxation bounds given $[R CS]$ conditional probability values	304
12.4	Description of variables in CPS data extract	305

12.5	Marginal table [$ACDGH$] from 8-way CPS table	306
12.6	Summary of difference between upper and lower bounds for small cell counts in the full 8-way CPS table under Model 1 and under Model 2	307
14.1	The Weather Dataset	338
14.2	Arbitrary partitioning of data between 2 sites	339
14.3	Vertical partitioning of data between 2 sites	340
15.1	Summarization of Attacks on Additive Perturbation	368
15.2	Summarization of Attacks on Matrix Multiplicative Perturbation	375
18.1	Notations and Definitions	441