## Contents

Preface v  
List of Figures xvii  
List of Tables xxi  

1  
An Introduction to Privacy-Preserving Data Mining  
Charu C. Aggarwal, Philip S. Yu  
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  
Charu C. Aggarwal, Philip S. Yu  
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
Josep Domingo-Ferrer
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
### Contents

4  
Measures of Anonymity  
Suresh Venkatasubramanian  
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  99  
4.5.1 New Directions  98  
References  99  

5  
$k$-Anonymous Data Mining: A Survey  
V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati  
5.1. Introduction  105  
5.2. $k$-Anonymity  107  
5.3. Algorithms for Enforcing $k$-Anonymity  110  
5.4. $k$-Anonymity Threats from Data Mining  117  
5.4.1 Association Rules  118  
5.4.2 Classification Mining  118  
5.5. $k$-Anonymity in Data Mining  120  
5.6. Anonymize-and-Mine  123  
5.7. Mine-and-Anonymize  126  
5.7.1 Enforcing $k$-Anonymity on Association Rules  126  
5.7.2 Enforcing $k$-Anonymity on Decision Trees  130  
5.8. Conclusions  133  
Acknowledgments  133  
References  134  

6  
A Survey of Randomization Methods for Privacy-Preserving Data Mining  
Charu C. Aggarwal, Philip S. Yu  
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

- **6.3.1 Privacy-Preserving Classification with Randomization**
- **6.3.2 Privacy-Preserving OLAP**
- **6.3.3 Collaborative Filtering**

### 6.4. The Privacy-Information Loss Tradeoff

### 6.5. Vulnerabilities of the Randomization Method

### 6.6. Randomization of Time Series Data Streams

### 6.7. Multiplicative Noise for Randomization

- **6.7.1 Vulnerabilities of Multiplicative Randomization**
- **6.7.2 Sketch Based Randomization**

### 6.8. Conclusions and Summary

References 154

---

### 7

#### A Survey of Multiplicative Perturbation for Privacy-Preserving Data Mining

**Keke Chen and Ling Liu**

- **7.1. Introduction**
  - 7.1.1 Data Privacy vs. Data Utility
  - 7.1.2 Outline
- **7.2. Definition of Multiplicative Perturbation**
  - 7.2.1 Notations
  - 7.2.2 Rotation Perturbation
  - 7.2.3 Projection Perturbation
  - 7.2.4 Sketch-based Approach
  - 7.2.5 Geometric Perturbation
- **7.3. Transformation Invariant Data Mining Models**
  - 7.3.1 Definition of Transformation Invariant Models
  - 7.3.2 Transformation-Invariant Classification Models
  - 7.3.3 Transformation-Invariant Clustering Models
- **7.4. Privacy Evaluation for Multiplicative Perturbation**
  - 7.4.1 A Conceptual Multidimensional Privacy Evaluation Model
  - 7.4.2 Variance of Difference as Column Privacy Metric
  - 7.4.3 Incorporating Attack Evaluation
  - 7.4.4 Other Metrics
- **7.5. Attack Resilient Multiplicative Perturbations**
  - 7.5.1 Naive Estimation to Rotation Perturbation
  - 7.5.2 ICA-Based Attacks
  - 7.5.3 Distance-Inference Attacks
  - 7.5.4 Attacks with More Prior Knowledge
  - 7.5.5 Finding Attack-Resilient Perturbations
- **7.6. Conclusion**

Acknowledgment 178

References 179

---

### 8

#### A Survey of Quantification of Privacy Preserving Data Mining Algorithms

**Elisa Bertino, Dan Lin and Wei Jiang**

- **8.1. Introduction**
- **8.2. Metrics for Quantifying Privacy Level**
  - 8.2.1 Data Privacy
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.2 Result Privacy</td>
<td>191</td>
</tr>
<tr>
<td>8.3. Metrics for Quantifying Hiding Failure</td>
<td>192</td>
</tr>
<tr>
<td>8.4. Metrics for Quantifying Data Quality</td>
<td>193</td>
</tr>
<tr>
<td>8.4.1 Quality of the Data Resulting from the PPDM Process</td>
<td>193</td>
</tr>
<tr>
<td>8.4.2 Quality of the Data Mining Results</td>
<td>198</td>
</tr>
<tr>
<td>8.5. Complexity Metrics</td>
<td>200</td>
</tr>
<tr>
<td>8.6. How to Select a Proper Metric</td>
<td>201</td>
</tr>
<tr>
<td>8.7. Conclusion and Research Directions</td>
<td>202</td>
</tr>
<tr>
<td>References</td>
<td>202</td>
</tr>
<tr>
<td>9 A Survey of Utility-based Privacy-Preserving Data Transformation Methods</td>
<td></td>
</tr>
<tr>
<td>Ming Hua and Jian Pei</td>
<td></td>
</tr>
<tr>
<td>9.1. Introduction</td>
<td>208</td>
</tr>
<tr>
<td>9.1.1 What is Utility-based Privacy Preservation?</td>
<td>209</td>
</tr>
<tr>
<td>9.2. Types of Utility-based Privacy Preservation Methods</td>
<td></td>
</tr>
<tr>
<td>9.2.1 Privacy Models</td>
<td>210</td>
</tr>
<tr>
<td>9.2.2 Utility Measures</td>
<td>212</td>
</tr>
<tr>
<td>9.2.3 Summary of the Utility-Based Privacy Preserving Methods</td>
<td>214</td>
</tr>
<tr>
<td>9.3. Utility-Based Anonymization Using Local Recoding</td>
<td>214</td>
</tr>
<tr>
<td>9.3.1 Global Recoding and Local Recoding</td>
<td>215</td>
</tr>
<tr>
<td>9.3.2 Utility Measure</td>
<td>216</td>
</tr>
<tr>
<td>9.3.3 Anonymization Methods</td>
<td>217</td>
</tr>
<tr>
<td>9.3.4 Summary and Discussion</td>
<td>219</td>
</tr>
<tr>
<td>9.4. The Utility-based Privacy Preserving Methods in Classification Problems</td>
<td>219</td>
</tr>
<tr>
<td>9.4.1 The Top-Down Specialization Method</td>
<td>220</td>
</tr>
<tr>
<td>9.4.2 The Progressive Disclosure Algorithm</td>
<td>224</td>
</tr>
<tr>
<td>9.4.3 Summary and Discussion</td>
<td>228</td>
</tr>
<tr>
<td>9.5. Anonymized Marginal: Injecting Utility into Anonymized Data Sets</td>
<td>228</td>
</tr>
<tr>
<td>9.5.1 Anonymized Marginal</td>
<td>229</td>
</tr>
<tr>
<td>9.5.2 Utility Measure</td>
<td>230</td>
</tr>
<tr>
<td>9.5.3 Injecting Utility Using Anonymized Marginals</td>
<td>231</td>
</tr>
<tr>
<td>9.5.4 Summary and Discussion</td>
<td>233</td>
</tr>
<tr>
<td>9.6. Summary</td>
<td>234</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>234</td>
</tr>
<tr>
<td>References</td>
<td>234</td>
</tr>
<tr>
<td>10 Mining Association Rules under Privacy Constraints</td>
<td>239</td>
</tr>
<tr>
<td>Jayant R. Haritsa</td>
<td></td>
</tr>
<tr>
<td>10.1. Introduction</td>
<td>239</td>
</tr>
<tr>
<td>10.2. Problem Framework</td>
<td>240</td>
</tr>
<tr>
<td>10.2.1 Database Model</td>
<td>240</td>
</tr>
<tr>
<td>10.2.2 Mining Objective</td>
<td>241</td>
</tr>
<tr>
<td>10.2.3 Privacy Mechanisms</td>
<td>241</td>
</tr>
<tr>
<td>10.2.4 Privacy Metric</td>
<td>243</td>
</tr>
<tr>
<td>10.2.5 Accuracy Metric</td>
<td>245</td>
</tr>
</tbody>
</table>
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
Vassilios S. Verykios and Aris Gkoulalas-Divanis
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
Stephen E. Fienberg and Aleksandra B. Slavkovic
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
Murat Kantarcioglu
13.1. Introduction 313
Contents

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
Jaideep Vaidya
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
Kun Liu, Chris Giannella, and Hillol Kargupta
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
Privacy-Preserving Data Mining
Models and Algorithms
Aggarwal, C.C.; Yu, P.S. (Eds.)
2008, XXII, 514 p., Hardcover