
Contents

1	Introduction	1
1.1	Motivations for Data Privacy	2
1.2	Privacy and Society	5
1.3	Terminology	6
1.3.1	The Framework	7
1.3.2	Anonymity and Unlinkability	8
1.3.3	Disclosure	10
1.3.4	Undetectability and Unobservability	13
1.3.5	Pseudonyms and Identity	14
1.3.6	Transparency	16
1.4	Privacy and Disclosure	17
1.5	Privacy by Design	17
	References	19
2	Machine and Statistical Learning	23
2.1	Classification of Techniques	24
2.2	Supervised Learning	25
2.2.1	Classification	25
2.2.2	Regression	26
2.2.3	Validation of Results: k-Fold Cross-Validation	27
2.3	Unsupervised Learning	28
2.3.1	Clustering	28
2.3.2	Association Rules	44
2.3.3	Expectation-Maximization Algorithm	51
	References	53
3	On the Classification of Protection Procedures	55
3.1	Dimensions	55
3.1.1	On Whose Privacy Is Being Sought	56
3.1.2	On the Computations to be Done	58
3.1.3	On the Number of Data Sources	60
3.1.4	Knowledge Intensive Data Privacy	61
3.1.5	Other Dimensions and Discussion	62
3.1.6	Summary	63

3.2	Respondent and Holder Privacy	64
3.3	Data-Driven Methods	65
3.4	Computation-Driven Methods.	66
3.4.1	Single Database: Differential Privacy	66
3.4.2	Multiple Databases: Cryptographic Approaches	69
3.4.3	Discussion.	70
3.5	Result-Driven Approaches	71
3.6	Tabular Data.	76
3.6.1	Cell Suppression	80
3.6.2	Controlled Tabular Adjustment	83
	References.	85
4	User's Privacy	89
4.1	User Privacy in Communications	90
4.1.1	Protecting the Identity of the User	90
4.1.2	Protecting the Data of the User.	95
4.2	User Privacy in Information Retrieval	97
4.2.1	Protecting the Identity of the User	97
4.2.2	Protecting the Query of the User.	98
4.3	Private Information Retrieval	101
4.3.1	Information-Theoretic PIR with k Databases.	101
4.3.2	Computational PIR	104
4.3.3	Other Contexts	108
	References.	108
5	Privacy Models and Disclosure Risk Measures	111
5.1	Definition and Controversies.	111
5.1.1	A Boolean or Measurable Condition.	114
5.2	Attribute Disclosure	115
5.2.1	Attribute Disclosure for a Numerical Variable	115
5.2.2	Attribute Disclosure for a Categorical Variable.	116
5.3	Identity Disclosure	118
5.3.1	An Scenario for Identity Disclosure	118
5.3.2	Measures for Identity Disclosure.	121
5.3.3	Uniqueness	122
5.3.4	Reidentification	123
5.3.5	The Worst-Case Scenario	125
5.4	Matching and Integration: A Database Based Approach.	126
5.4.1	Heterogenous Distributed Databases	127
5.4.2	Data Integration	127
5.4.3	Schema Matching	131
5.4.4	Data Matching	133
5.4.5	Preprocessing	134
5.4.6	Indexing and Blocking	135

5.4.7	Record Pair Comparison: Distances and Similarities	136
5.4.8	Classification of Record Pairs	140
5.5	Probabilistic Record Linkage	142
5.5.1	Alternative Expressions for Decision Rules	152
5.5.2	Computation of $R_p(a, b)$	156
5.5.3	Estimation of the Probabilities	158
5.5.4	Extensions for Computing Probabilities	160
5.5.5	Final Notes	163
5.6	Distance-Based Record Linkage	163
5.6.1	Weighted Distances	164
5.6.2	Distance and Normalization	167
5.6.3	Parameter Determination for Record Linkage	168
5.7	Record Linkage Without Common Variables	171
5.8	k -Anonymity and Other Boolean Conditions for Identity Disclosure	172
5.8.1	k -Anonymity and Anonymity Sets: k -Confusion	173
5.8.2	k -Anonymity and Attribute Disclosure: Attacks	176
5.8.3	Other Related Approaches to k -Anonymity	178
5.9	Discussion on Record Linkage	178
5.9.1	Formalization of Reidentification Algorithms	178
5.9.2	Comparison of Record Linkage Algorithms	179
5.9.3	Disclosure Risk and Big Data	180
5.9.4	Some Guidelines and Research Issues	182
	References	183
6	Masking Methods	191
6.1	Perturbative Methods	193
6.1.1	Data and Rank Swapping	193
6.1.2	Microaggregation	200
6.1.3	Additive and Multiplicative Noise	212
6.1.4	PRAM	214
6.1.5	Lossy Compression and Other Transform-Based Method	217
6.2	Non-perturbative Methods	219
6.2.1	Generalization and Recoding	219
6.2.2	Suppression	220
6.3	Synthetic Data Generators	220
6.4	Masking Methods and k -anonymity	223
6.4.1	Mondrian	224
6.4.2	Algorithms for k -anonymity: Variants and Big Data	225
6.5	Data Protection Procedures for Constrained Data	226
6.5.1	Types of Constraints	227
6.6	Masking Methods and Big Data	230
	References	231

7	Information Loss: Evaluation and Measures	239
7.1	Generic Versus Specific Information Loss	239
7.2	Information Loss Measures	240
7.3	Generic Information Loss Measures	242
7.3.1	Numerical Data	242
7.3.2	Categorical Data	245
7.4	Specific Information Loss	248
7.4.1	Classification-Based Information Loss	248
7.4.2	Regression-Based Information Loss	249
7.4.3	Clustering-Based Information Loss	249
7.5	Information Loss and Big Data	250
	References	251
8	Selection of Masking Methods	255
8.1	Aggregation: A Score	255
8.2	Visualization: R-U Maps	256
8.3	Optimization and Postmasking	257
	References	258
9	Conclusions	259
	Reference	260
	Index	261

Data Privacy: Foundations, New Developments and the
Big Data Challenge

Torra, V.

2017, XIV, 269 p. 22 illus., Hardcover

ISBN: 978-3-319-57356-4