Contents

1 Introduction 1
- 1.1 Motivation 1
- 1.2 Challenges 4
- 1.3 Focus of the Book 5
- 1.4 Organization of the Book 6

2 Probabilistic Ranking Queries on Uncertain Data 9
- 2.1 Basic Uncertain Data Models 9
  - 2.1.1 Uncertain Object Model 9
  - 2.1.2 Probabilistic Database Model 11
  - 2.1.3 Converting Between the Uncertain Object Model and the Probabilistic Database Model 12
- 2.2 Basic Ranking Queries on Uncertain Data 12
  - 2.2.1 Ranking Instances in An Uncertain Object 13
  - 2.2.2 Ranking Uncertain Instances in Multiple Uncertain Objects . 19
  - 2.2.3 Ranking Uncertain Objects 22
- 2.3 Extended Uncertain Data Models and Ranking Queries 22
  - 2.3.1 Uncertain Data Stream Model 22
  - 2.3.2 Probabilistic Linkage Model 25
  - 2.3.3 Uncertain Road Network 27
- 2.4 Summary 31

3 Related Work 33
- 3.1 Uncertain Data Processing 33
  - 3.1.1 Uncertain Data Models and Systems 33
  - 3.1.2 Probabilistic Queries on Uncertain Data 34
  - 3.1.3 Indexing Uncertain Data 35
- 3.2 Ranking (Top-k) Queries 35
  - 3.2.1 Distributed Top-k Query Processing 36
- 3.3 Top-k Typicality Queries 36
  - 3.3.1 Typicality in Psychology and Cognitive Science 36

3.3.2 The (Discrete) k-Median Problem 37
3.3.3 Clustering Analysis 38
3.3.4 Other Related Models 39
3.4 Probabilistic Ranking Queries 40
3.4.1 Top-k Queries on Uncertain Data 40
3.4.2 Poisson Approximation 42
3.5 Uncertain Streams 43
3.5.1 Continuous Queries on Probabilistic Streams 43
3.5.2 Continuous Ranking and Quantile Queries on Data Streams . 44
3.5.3 Continuous Sensor Stream Monitoring 45
3.6 Probabilistic Linkage Queries 46
3.6.1 Record Linkage 46
3.6.2 Probabilistic Graphical Models 47
3.7 Probabilistic Path Queries 48
3.7.1 Path Queries on Probabilistic Graphs 48
3.7.2 Path Queries on Certain Traffic Networks 49

4 Top-k Typicality Queries on Uncertain Data 51
4.1 Answering Simple Typicality Queries 52
4.1.1 Likelihood Computation 52
4.1.2 An Exact Algorithm and Complexity 53
4.1.3 A Randomized Tournament Algorithm 54
4.2 Local Typicality Approximation 56
4.2.1 Locality of Typicality Approximation 56
4.2.2 DLTA: Direct Local Typicality Approximation Using VP-trees 59
4.2.3 LT3: Local Typicality Approximation Using Tournaments . . 61
4.3 Answering Discriminative Typicality Queries 65
4.3.1 A Randomized Tournament Algorithm 65
4.3.2 Local Typicality Approximation 66
4.4 Answering Representative Typicality Queries 69
4.4.1 An Exact Algorithm and Complexity 69
4.4.2 A Randomized Tournament Method 70
4.4.3 Local Typicality Approximation Methods 70
4.5 Empirical Evaluation 74
4.5.1 Typicality Queries on Real Data Sets 75
4.5.2 Approximation Quality 80
4.5.3 Sensitivity to Parameters and Noise 85
4.5.4 Efficiency and Scalability 86
4.6 Summary 87

5 Probabilistic Ranking Queries on Uncertain Data 89
5.1 Top-k Probability Computation 90
5.1.1 The Dominant Set Property 90
5.1.2 The Basic Case: Independent Tuples 91

5.1.3 Handling Generation Rules 92
5.2 Exact Query Answering Methods 94
5.2.1 Query Answering Framework 94
5.2.2 Scan Reduction by Prefix Sharing 95
5.2.3 Pruning Techniques 100
5.3 A Sampling Method 102
5.4 A Poisson Approximation Based Method 104
5.4.1 Distribution of Top-*k* Probabilities 104
5.4.2 A General Stopping Condition 105
5.4.3 A Poisson Approximation Based Method 106
5.5 Online Query Answering 107
5.5.1 The *PRist* Index 107
5.5.2 Query Evaluation based on *PRist* 110
5.5.3 *PRist*+ and a Fast Construction Algorithm 115
5.6 Experimental Results 117
5.6.1 Results on IIP Iceberg Database 117
5.6.2 Results on Synthetic Data Sets 120
5.7 Summary 127

6 Continuous Ranking Queries on Uncertain Streams 129
6.1 Exact Algorithms 129
6.1.1 Top-k Probabilities in a Sliding Window 130
6.1.2 Sharing between Sliding Windows 133
6.2 A Sampling Method 138
6.3 Space Efficient Methods 138
6.3.1 Top-*k* Probabilities and Quantiles 139
6.3.2 Approximate Quantile Summaries 142
6.3.3 Space Efficient Algorithms using Quantiles 144
6.4 Experimental Results 145
6.4.1 Results on Real Data Sets 145
6.4.2 Synthetic Data Sets 146
6.4.3 Efficiency and Approximation Quality 147
6.4.4 Scalability 150
6.5 Summary 150

7 Ranking Queries on Probabilistic Linkages 151
7.1 Review: the Probabilistic Linkage Model 151
7.2 Linkage Compatibility 152
7.2.1 Dependencies among Linkages 152
7.2.2 Probabilistic Mutual Exclusion Graphs 153
7.2.3 Compatibility of Linkages 155
7.2.4 Resolving Incompatibility 157
7.2.5 Deriving All Possible Worlds 158
7.3 Ranking Queries on Probabilistic Linkages 160
7.3.1 Predicate Processing 161

7.3.2 Dominant Subgraphs . . . 162
7.3.3 Vertex Compression . . . 163
7.3.4 Subgraph Probabilities . . . 164
7.4 Tree Recurrence: Subgraph Probability Calculation . . . 165
7.4.1 A Chain of Cliques . . . 165
7.4.2 A Tree of Cliques . . . 167
7.4.3 Integrating Multiple Components . . . 168
7.5 Exact Query Answering Algorithms . . . 169
7.5.1 An Exact Algorithm . . . 169
7.5.2 Reusing Intermediate Results . . . 169
7.5.3 Pruning Techniques . . . 171
7.6 Extensions to Aggregate Queries . . . 172
7.6.1 Aggregate Queries on Probabilistic Linkages . . . 172
7.6.2 Count, Sum and Average Queries . . . 173
7.6.3 Min and Max Queries . . . 177
7.7 Empirical Evaluation . . . 178
7.7.1 Results on Real Data Sets . . . 178
7.7.2 Results on Synthetic Data Sets . . . 182
7.8 Summary . . . 184

8 Probabilistic Path Queries on Road Networks . . . 185
8.1 Probability Calculation . . . 186
8.1.1 Exact l-Weight Probability Calculation . . . 186
8.1.2 Approximating l-Weight Probabilities . . . 188
8.1.3 Estimating l-Weight Probabilities . . . 191
8.1.4 A Depth First Search Method . . . 191
8.2 P*: A Best First Search Method . . . 192
8.2.1 The P* Algorithm . . . 193
8.2.2 Heuristic Estimates . . . 194
8.3 A Hierarchical Index for P* . . . 199
8.3.1 HP-Tree Index . . . 199
8.3.2 Approximating Min-Value Estimates . . . 200
8.3.3 Approximating Stochastic Estimates . . . 200
8.4 Experimental Results . . . 201
8.4.1 Simulation Setup . . . 202
8.4.2 Efficiency and Memory Usage . . . 204
8.4.3 Approximation Quality and Scalability . . . 205
8.5 Summary . . . 206

9 Conclusions . . . 207
9.1 Summary of the Book . . . 207
9.2 Future Directions: Possible Direct Extensions . . . 209
9.2.1 Top-k typicality queries for uncertain object . . . 210
9.2.2 Top-k queries on probabilistic databases . . . 210
9.2.3 Probabilistic ranking queries . . . 210

9.2.4 Probabilistic path queries on road networks 212
9.3 Future Directions: Enriching Data Types and Queries 213
9.3.1 Handling Complex Uncertain Data and Data Correlations . . 213
9.3.2 Answering More Types of Ranking and Preference Queries on Uncertain Data 213
References 215



http://www.springer.com/978-1-4419-9379-3

Ranking Queries on Uncertain Data
Hua, M.; Pei, J.
2011, XVI, 224 p., Hardcover
ISBN: 978-1-4419-9379-3