Table of Contents

Preface ix

Foreword xiii

1. Introduction 1
- What Are Graphs? 2
- What Are Graph Analytics and Algorithms? 3
- Graph Processing, Databases, Queries, and Algorithms 6
  - OLTP and OLAP 7
- Why Should We Care About Graph Algorithms? 8
- Graph Analytics Use Cases 12
- Conclusion 14

2. Graph Theory and Concepts 15
- Terminology 15
- Graph Types and Structures 16
  - Random, Small-World, Scale-Free Structures 17
- Flavors of Graphs 18
  - Connected Versus Disconnected Graphs 19
  - Unweighted Graphs Versus Weighted Graphs 19
  - Undirected Graphs Versus Directed Graphs 21
  - Acyclic Graphs Versus Cyclic Graphs 22
  - Sparse Graphs Versus Dense Graphs 23
  - Monopartite, Bipartite, and k-Partite Graphs 24
- Types of Graph Algorithms 27
  - Pathfinding 27
  - Centrality 27
  - Community Detection 27

Summary 28

3. **Graph Platforms and Processing. 29**
Graph Platform and Processing Considerations 29
Platform Considerations 29
Processing Considerations 30
Representative Platforms 31
Selecting Our Platform 31
Apache Spark 32
Neo4j Graph Platform 34
Summary 37

4. **Pathfinding and Graph Search Algorithms. 39**
Example Data: The Transport Graph 41
Importing the Data into Apache Spark 43
Importing the Data into Neo4j 44
Breadth First Search 45
Breadth First Search with Apache Spark 46
Depth First Search 48
Shortest Path 49
When Should I Use Shortest Path? 50
Shortest Path with Neo4j 51
Shortest Path (Weighted) with Neo4j 53
Shortest Path (Weighted) with Apache Spark 54
Shortest Path Variation: A* 56
Shortest Path Variation: Yen's k-Shortest Paths 58
All Pairs Shortest Path 60
A Closer Look at All Pairs Shortest Path 60
When Should I Use All Pairs Shortest Path? 62
All Pairs Shortest Path with Apache Spark 62
All Pairs Shortest Path with Neo4j 63
Single Source Shortest Path 65
When Should I Use Single Source Shortest Path? 67
Single Source Shortest Path with Apache Spark 67
Single Source Shortest Path with Neo4j 69
Minimum Spanning Tree 70
When Should I Use Minimum Spanning Tree? 71
Minimum Spanning Tree with Neo4j 72
Random Walk 73
When Should I Use Random Walk? 74
Random Walk with Neo4j 74
Summary 75

5. Centrality Algorithms. 77
Example Graph Data: The Social Graph 79
Importing the Data into Apache Spark 80
Importing the Data into Neo4j 81
Degree Centrality 81
Reach 81
When Should I Use Degree Centrality? 82
Degree Centrality with Apache Spark 83
Closeness Centrality 84
When Should I Use Closeness Centrality? 85
Closeness Centrality with Apache Spark 86
Closeness Centrality with Neo4j 88
Closeness Centrality Variation: Wasserman and Faust 89
Closeness Centrality Variation: Harmonic Centrality 91
Betweenness Centrality 92
When Should I Use Betweenness Centrality? 94
Betweenness Centrality with Neo4j 95
Betweenness Centrality Variation: Randomized-Approximate Brandes 98
PageRank 99
Influence 99
The PageRank Formula 100
Iteration, Random Surfers, and Rank Sinks 102
When Should I Use PageRank? 103
PageRank with Apache Spark 103
PageRank with Neo4j 105
PageRank Variation: Personalized PageRank 107
Summary 108

6. Community Detection Algorithms. 109
Example Graph Data: The Software Dependency Graph 112
Importing the Data into Apache Spark 114
Importing the Data into Neo4j 114
Triangle Count and Clustering Coefficient 114
Local Clustering Coefficient 115
Global Clustering Coefficient 116
When Should I Use Triangle Count and Clustering Coefficient? 116
Triangle Count with Apache Spark 117
Triangles with Neo4j 117
Local Clustering Coefficient with Neo4j 118
Strongly Connected Components 119
When Should I Use Strongly Connected Components? 120
Strongly Connected Components with Apache Spark 120

Strongly Connected Components with Neo4j 122
Connected Components 124
When Should I Use Connected Components? 124
Connected Components with Apache Spark 125
Connected Components with Neo4j 126
Label Propagation 127
Semi-Supervised Learning and Seed Labels 129
When Should I Use Label Propagation? 129
Label Propagation with Apache Spark 130
Label Propagation with Neo4j 131
Louvain Modularity 133
When Should I Use Louvain? 137
Louvain with Neo4j 138
Validating Communities 143
Summary 143

7. Graph Algorithms in Practice. 145
Analyzing Yelp Data with Neo4j 145
Yelp Social Network 146
Data Import 147
Graph Model 147
A Quick Overview of the Yelp Data 148
Trip Planning App 152
Travel Business Consulting 157
Finding Similar Categories 162
Analyzing Airline Flight Data with Apache Spark 166
Exploratory Analysis 168
Popular Airports 168
Delays from ORD 170
Bad Day at SFO 172
Interconnected Airports by Airline 174
Summary 181

8. Using Graph Algorithms to Enhance Machine Learning. 183
Machine Learning and the Importance of Context 183
Graphs, Context, and Accuracy 184
Connected Feature Extraction and Selection 185
Graphy Features 187
Graph Algorithm Features 188
Graphs and Machine Learning in Practice: Link Prediction 190
Tools and Data 190
Importing the Data into Neo4j 192

The Coauthorship Graph 193
Creating Balanced Training and Testing Datasets 194
How We Predict Missing Links 199
Creating a Machine Learning Pipeline 200
Predicting Links: Basic Graph Features 201
Predicting Links: Triangles and the Clustering Coefficient 214
Predicting Links: Community Detection 218
Summary 224
Wrapping Things Up 224

A. Additional Information and Resources. 225

Index. 231