Index

A

A* algorithm
  Shortest Path, 56-58
  with Neo4j, 57
actual density, 24
acyclic graphs, 18
  cyclic graphs vs., 22-23
  trees and, 23
aggregateMessages, 54, 67, 86
airline flight data
  analyzing with Apache Spark, 166-181
  delays from ORD, 170-172
  exploratory analysis, 168
  fog-related delays from SFO, 172-174
  interconnected airports by airline, 174-181
  popular airports, 168-170
Alexa, 185
algorithm-centric processing, 30
All Pairs Shortest Path (APSP) algorithm, 40, 60-65
  sequence of operations, 60-62
  when to use, 62
  with Apache Spark, 62
  with Neo4j, 63-65
Amazon, 185
anti-money laundering (AML) analysis, xiv
Apache Spark
  about, 32-34
  All Pairs Shortest Path algorithm with, 62
  analyzing airline flight data with, 166-181
    (see also airline flight data)
  Breadth First Search with, 46
  Closeness Centrality with, 86-88
  Connected Components with, 125-126
  Degree Centrality with, 83
  importing social graph data into, 80
  importing software dependency graph data into, 114
  importing transport dataset into, 43
  installing, 34
  Label Propagation with, 130
  online resources, 228
  PageRank with, 103-105
  personalized PageRank with, 107
  Shortest Path algorithm (weighted), 54-56
  Single Source Shortest Path algorithm with, 67-69
  Spark Graph project, 33
  Strongly Connected Components with, 120
  Triangle Count with, 117
  when to use, 31
Approximate Betweenness Centrality, 160
artificial intelligence (AI), 183
average degree, 81
average shortest path, 27
Awesome Procedures on Cypher (APOC) library, 35, 148, 192, 228

B

Bacon number, 50
Barabási, Albert-László, 11
Betweenness Centrality algorithm, 78, 92-99
  bridges and control points, 93
  calculating, 93
  when to use, 94
  with Neo4j, 95-97
  with Yelp dataset, 159-162
binary classification, 200

bipartite graphs, 18, 25
Boruvka, Otakar, 70
Breadth First Search (BFS), 45-47
bridges, 93
Bridges of Königsberg problem, 2, 49
bulk data import, Neo4j Import tool for, 225-227
Bulk Synchronous Parallel (BSP), 30

C
cancer research, xiv
Centrality algorithms, 27, 77-108
  Betweenness Centrality, 92-99
  Closeness Centrality, 84-92
  Degree Centrality, 81-84
  overview, 78
  Randomized-Approximate Brandes, 98
  social graph data for, 79-81
Chicago O'Hare International Airport (ORD), data in delays from, 170-172
citation networks, 124
Clauset, A., 5
clique, 23
Closeness Centrality algorithm, 78, 84-92
  Harmonic Centrality variation, 91
  Wasserman Faust variation, 89-91
  when to use, 85
  with Apache Spark, 86-88
  with Neo4j, 88
Clustering Coefficient algorithm, 109
  (see also Triangle Count and Clustering Coefficient algorithms)
clusters, defined, 19
Colorado Index of Complex Networks (ICON), 228
community detection algorithms, 27, 109-143
  Connected Components, 124-127
  for link prediction, 218-223
  Label Propagation algorithm, 127-133
  Louvain Modularity, 133-143
  software dependency graph data for, 112-114
  Strongly Connected Components, 119-124
  Triangle Count and Clustering Coefficient, 114-119
  validating communities, 143
complete graph, 23
components, defined, 19
Configuration that Outperforms a Single Thread (COST), 29
Connected Components algorithm, 109, 124-127
  when to use, 124
  with Apache Spark, 125-126
  with Neo4j, 126
connected graphs, 19
context, xiv, 183
costs, 50, 57
cycles, 22
cyclic graphs, 18
Cypher, 33, 166, 225
Cypher for Apache Spark (CAPS), 33

D
D'Orazio, Francesco, 10
damping/dampening factor, 102, 104
DataFrame, 32
datasets, sources for, 228
deduplication, 124
Degree Centrality algorithm, 78, 81-84
  reach of a node, 81
  when to use, 82
  with airport data, 168-170
  with Apache Spark, 83
degree distribution, 81
degree of a node, 81
delta-stepping algorithm, 69
dense graphs, 18, 23
density of relationships, 111
Depth First Search (DFS), 48-49, 120
diameter of a graph, 27
Dijkstra, Edsger, 50
Dijkstra's algorithm (see Shortest Path algorithm)
directed acyclic graph (DAG), 22
directed graphs, 18, 21
directional relationships, xiii
disconnected graphs, 19
distance (term), 50

E
Eguíluz, Víctor M., 4
entity relationship diagram (ERD), xiii
Erdös, Paul, 51
Euler, Leonhard, 2
Eulerian path, 49

F

Facebook, 184
Faust, Katherine, 89
feature extraction, 186
feature importance, 210
feature selection, 186
feature vectors, 186
features
  connected feature extraction/selection, 185-189
  graph algorithm features, 188
  graphy, 187
Fischer, Michael J., 124
Fleurquin, Pablo, 4
foodweb, 5
Freeman, Linton C., 81, 93

G

Galler, Bernard A., 124
Girvan–Newman (GN) benchmark, 143
global clustering coefficient, 116
global patterns, 6
Google
  PageRank, 99
  Pregel, 30
Grandjean, Martin, 4
graph algorithms (generally)
  about, 3-5
  centrality, 27
  community detection, 27
  defined, 3
    (see also specific algorithms)
  importance of, 8-12
  in practice, 145-181
  pathfinding, 27
  types of, 27
graph analytics
  about, 3-5
  defined, 3
  use cases, 12
graph compute engines, 31
graph databases, 31
graph embedding, defined, 186
graph global, 6, 186
graph local, 6, 186
graph platforms
  Apache Spark, 32-34
  Neo4j, 34-37
  platform considerations, 29
  representative platforms, 31-37
  selecting a platform, 31
graph processing, 6-8, 30
Graph search algorithms, 39-49
  defined, 39
Graph Search algorithms, 40
  Breadth First Search, 45-47
  Depth First Search, 48-49
  transport graph data for, 41-44
graph theory, 15-28
  about, 15-28
  origins of, 2
  terminology, 15
  types and structures, 16
graph traversal algorithms
  Breadth First Search, 45-47
  Depth First Search, 48-49
graph-centric processing, 30
GraphFrames, 32, 103, 114
graphs (generally)
  about, 2
  acyclic vs. cyclic, 18, 22-23
  bipartite, 18, 25
  common attributes, 18
  connected vs. disconnected, 19
  flavors of, 18-25
  k-partite, 18, 25
  monopartite, 18, 24
  sparse vs. dense, 23
  undirected vs. directed, 18, 21
  unweighted vs. weighted, 18, 19
graphy datasets, 228
graphy features, 187

H

Hamiltonian path, 49
Harmonic Centrality closeness algorithm, 91
Hart, Peter, 57
hop (term), 21, 50
hybrid transactional and analytical processing (HTAP), 7

I

impurity, 211
in-links, 21
influence, 99
islands, 19

K

k-partite graphs, 18, 25
Koblenz Network Collection (KONECT), 228
Königsberg Bridges problem, 2, 49

L

Label Propagation algorithm (LPA), 109, 127-133
  pull method, 128
  push method, 127
  seed labels, 129
  semi-supervised learning, 129
  when to use, 129
  with Apache Spark, 130
  with Neo4j, 131-133
  with Yelp dataset, 163-165
label, defined, 15
labeled property graph model, 15
Lancichinetti–Fortunato–Radicchi (LFR) benchmark, 143
landmarks, 62
Latora, V., 91
leaf nodes, 22
Lee, C. Y., 45
link prediction, 190-223
  balancing/splitting data for training/testing, 197-199
  basic graph features for, 201-213
  coauthorship graph, 193
  community detection, 218-223
  creating balanced training and testing datasets, 194-199
  creating machine learning pipeline, 200
  defined, 189
  importing data into Neo4j, 192
  predicting missing links, 199
  tools and data, 190-192
  Triangles and Clustering Coefficient, 214-218
literature-based discovery (LBD), xiv
local clustering coefficient, 115, 118
Louvain Modularity algorithm, 109, 133-143
  for link prediction, 219-221
  quality-based grouping via modularity, 134-137
  when to use, 137
  with Neo4j, 138-143

M

machine learning (ML)
  connected feature extraction/selection, 185-189
  graph embeddings, 186
  graphs, context, and accuracy, 184
  importance of context in, 183
  link prediction, 183
Marchiori, M., 91
marketing campaigns, xiv
matplotlib, 148
maximum density, 24
Minimum Spanning Tree algorithm, 40, 70-73
  when to use, 71
  with Neo4j, 72
modularity, 134
  (see also Louvain Modularity algorithm)
  calculating, 134-135
  quality-based grouping and, 134-137
money laundering, xiv
monopartite graphs, 18, 24
Moore, C., 5
Moore, Edward F., 45
multigraph, 17

N

negative weights, 51
Neo4j
  A* algorithm with, 57
  All Pairs Shortest Path algorithm with, 63-65
  analyzing Yelp data with, 145-166
    (see also Yelp dataset)
  Betweenness Centrality with, 95-97
  Closeness Centrality with, 88
  Connected Components with, 126
  importing Citation Network Dataset into, 192
  importing social graph data into, 81
  importing software dependency graph data into, 114
  importing transport dataset into, 44
  Label Propagation with, 131-133
  local clustering coefficient with, 118
  Louvain Modularity with, 138-143
  Minimum Spanning Tree algorithm with, 72
  online resources, 228
  PageRank with, 105-107
  Random Walk algorithm with, 74

Randomized-Approximate Brandes with, 98
Shortest Path algorithm (unweighted), 51-53
Shortest Path algorithm (weighted), 53
Single Source Shortest Path algorithm with, 69
Strongly Connected Components with, 122-124
Triangles with, 117
when to use, 32
Yen's k-Shortest Paths algorithm, 58
Neo4j Algorithms library
Shortest Path (unweighted), 51-53
Shortest Path (weighted), 53
Neo4j Desktop, 36
Neo4j Graph Platform, 34-37
Neo4j Import tool, 225-227
Network Science, 4
networks
graph as representation of, 3
types and structures, 16
Newman, M. E. J., 5
Nilsson, Nils, 57
node-centric processing, 30
nodes
Centrality and, 77
defined, 2

O

online analytical processing (OLAP), 7
online learning, 229
online transaction processing (OLTP), 7
out-links, 21

P

Page, Larry, 99
PageRank, 78, 99-108
and influence, 99
convergence implementation, 105
formula for, 100
iteration/random surfers/rank sinks, 102
Personalized PageRank variant, 107
when to use, 103
with Apache Spark, 103-105
with fixed number of iterations, 104
with Neo4j, 105-107
with Yelp dataset, 154-157
pandas library, 148
Pareto distribution, 11
path, defined, 16
Pathfinding algorithms, 27, 39-44
All Pairs Shortest Path, 60-65
Minimum Spanning Tree algorithm, 70-73
Random Walk algorithm, 73-75
Shortest Path, 49-60
Single Source Shortest Path, 65-70
transport graph data for, 41-44
weighted graphs and, 20
Pearson, Karl, 73
Personalized PageRank (PPR), 103, 107
pivotal nodes, 93
power law, 11
preferential attachment, 9
Pregel, 30
Prim's algorithm, 70
product recommendation engines, 185
properties, defined, 16
pseudograph, 17
pyspark REPL, 34

Q

quality-based grouping, 134-137

R

Raghavan, Usha Nandini, 127
Ramasco, José J., 4
random forest, 200, 213
random network, 18
Random Walk algorithm, 40, 73-75
when to use, 74
with Neo4j, 74
Randomized-Approximate Brandes (RA-Brandes) centrality algorithm, 98
rank sink, 102
Raphael, Bertram, 57
reach of a node, 81
Reif, Jennifer, 37
relationship type, 16
relationship-centric processing, 30
relationships (term), 1, 2

S

San Francisco International Airport (SFO), data in fog-related delays from, 172-174
scale-free network, 18
scaling law (power law), 11
search engines, xiv

seed labels, 129
semi-supervised learning, 129
Seven Bridges of Königsberg problem, 2, 49
Shortest Path algorithm, 40, 49-60
  A* algorithm, 56-58
  when to use, 50
  with Apache Spark (weighted), 54-56
  with Neo4j (unweighted), 51-53
  with Neo4j (weighted), 53
  Yen's k-Shortest Paths variation, 58
simple graph, 16
Single Source Shortest Path (SSSP) algorithm, 40, 65-70
  with Apache Spark, 67-69
  with Neo4j, 69
small-world network, 18
social graph data
  for Centrality algorithms, 79-81
  importing into Apache Spark, 80
  importing into Neo4j, 81
social network analysis, 116
software dependency graph data, 112-114
  importing into Apache Spark, 114
  importing into Neo4j, 114
spanning trees, 23
Spark Graph project, 33
sparse graphs, 18, 23
Stanford Network Analysis Project (SNAP), 228
strict graph, 16
Strogatz, Steven, 9
Strongly Connected Components (SCC) algorithm, 109, 119-124
  when to use, 120
  with airport data, 175-180
  with Apache Spark, 120
  with Neo4j, 122-124
structural hole, 119
subgraph, defined, 16

T
teleportation, 102
testing datasets, 194-199
training datasets, 194-199
training, online resources for, 229
transitive relationships, xiii
translytics, 7
transport datasets, 41-44
  importing into Apache Spark, 43
  importing into Neo4j, 44
Traveling Salesman Problem (TSP), 49
traversal-centric processing, 30
trees, 23
Trémaux, Charles Pierre, 48
Triangle Count and Clustering Coefficient algorithms, 109, 114-119
  for link prediction (machine learning example), 214-218
  global clustering coefficient, 116
  local clustering coefficient, 115
  local clustering coefficient with Neo4j, 118
  Triangle Count with Apache Spark, 117
  Triangles with Neo4j, 117
  when to use, 116
trip planning app, 152-159
Twitter
  Label Propagation, 129
  Personalized PageRank, 103

U
undirected graphs, 18, 21
Union Find, 124
unweighted graphs, 18, 19
unweighted shortest paths, 51-53

V
vertices, 2
  (see also nodes)

W
Wasserman Faust closeness algorithm, 89-91
Wasserman, Stanley, 89
Weakly Connected Components, 124
weight (term), 50
weighted graphs, 18, 19
Weighted Shortest Paths
  with Apache Spark, 54-56
  with Neo4j, 53
weightProperty, 107

Y
Yelp dataset
  analyzing with Neo4j, 145-166
  Bellagio cross-promotion, 159-162
  finding influential hotel reviewers, 154-159
  finding similar categories, 162-166
  graph model, 147
  importing into Neo4j, 147

Neo4j Import tool for, 225-227
overview, 148-151
social network, 146
travel business consulting, 157-159
trip planning app, 152-159
Yen's k-Shortest Paths algorithm, 58
Yen, Jin Y., 58