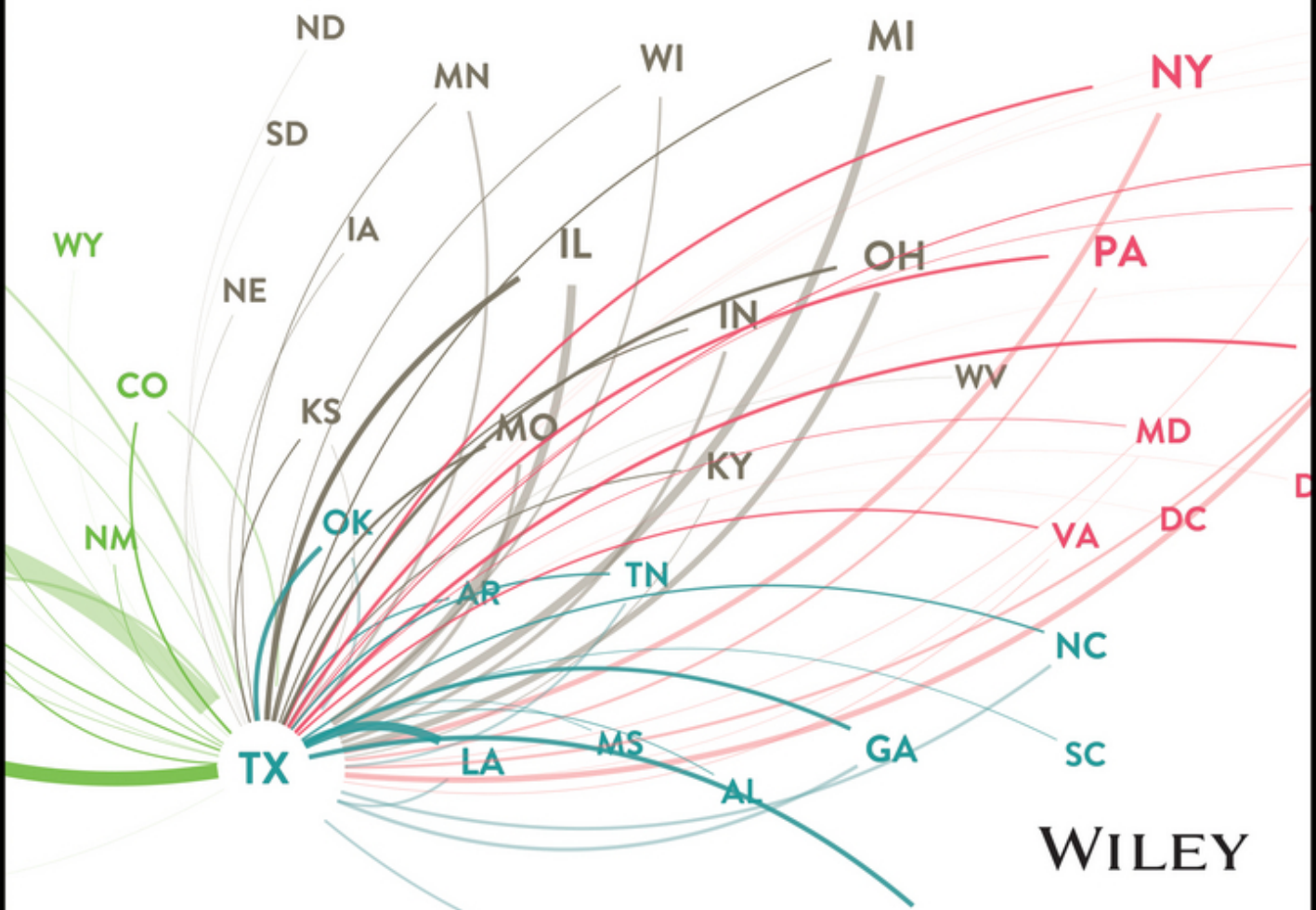


# Graph Analysis and Visualization

Discovering Business Opportunity in Linked Data

Richard Brath and David Jonker



WILEY



# Graph Analysis *and* Visualization



# Graph Analysis *and* Visualization

DISCOVERING BUSINESS OPPORTUNITY IN LINKED DATA

Richard Brath | David Jonker

WILEY

## **Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data**

Published by  
John Wiley & Sons, Inc.  
10475 Crosspoint Boulevard  
Indianapolis, IN 46256  
[www.wiley.com](http://www.wiley.com)

Copyright © 2015 by John Wiley & Sons, Inc., Indianapolis, Indiana  
Published simultaneously in Canada

ISBN: 978-1-118-84584-4  
ISBN: 978-1-118-84569-1 (ebk)  
ISBN: 978-1-118-84587-5 (ebk)

Manufactured in the United States of America  
10 9 8 7 6 5 4 3 2 1

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Limit of Liability/Disclaimer of Warranty:** The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

**Library of Congress Control Number:** 2014951021

**Trademarks:** Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

To Bayla, Abe, and Hana, who provide endless support for all my endeavors.

—*Richard Brath*

To Heather, Micah, Avril, and Naomi for their love and sacrifice in the making of this book. To Chris White for his vision and support in striving to put better tools in the hands of those who need them most.

—*David Jonker*





# ABOUT THE AUTHORS

**Richard Brath** is actively involved in the research, design, and development of data visualization and visual analytics for both research and commercial applications. His solutions range from rich interactive visualizations for mobile devices to large multi-touch, multi-screen installations, and web-based analytical visualizations for business applications. Brath's visualizations are used by hundreds of thousands of people every day in applications as diverse as trading, professional sports, and broadcast television.

**David Jonker** is a co-founder and Senior Partner of Uncharted (formerly Oculus Info Inc). He is a designer and developer of visual analytics tools and platforms for web-based, distributed, and mobile use. His work over the past two decades includes visualization systems and content for the NASDAQ MarketSite real-time broadcast center in Times Square. He is currently a lead on the DARPA XDATA program. Jonker and Brath are business partners and regular presenters and publishers of work in leading industry and research forums.



# ABOUT THE TECHNICAL EDITORS

**Scott Langevin** is a director and research scientist at Uncharted, with more than 12 years of industry and academic experience. He holds a PhD in computer science from the University of South Carolina, and has a background in machine learning, service-oriented computing, and software engineering. Langevin's research interests are in probabilistic graphical modeling, large-scale visual analytics, and adaptive user interfaces.

**Peter MacMurchy** has been a professional software developer for more than 15 years, focusing on UX, UI, and interactive data-visualization tools. He acquired a keen interest in information visualization from coursework while studying computer graphics for his master of science degree in computer science at the University of Calgary. Since then, he's continued to develop visualization and interactive software for finance, film, energy, and other industries.



# CREDITS

**Executive Editor**

Robert Elliott

**Project Editor**

Kevin Shafer

**Technical Editors**

Scott Langevin

Peter MacMurphy

**Production Editor**

Rebecca Anderson

**Copy Editor**

Kim Cofer

**Manager of Content Development  
and Assembly**

Mary Beth Wakefield

**Marketing Director**

David Mayhew

**Marketing Manager**

Carrie Sherrill

**Professional Technology and  
Strategy Director**

Barry Pruett

**Business Manager**

Amy Knies

**Associate Publisher**

Jim Minatel

**Production Manager**

Kathleen Wisor

**Project Coordinator, Cover**

Patrick Redmond

**Composer**

Maureen Forsys,

Happenstance Type-O-Rama

**Proofreader**

Kim Wimpsett

**Indexer**

Johnna VanHoose

**Cover Designer**

Wiley

**Cover Image**

Courtesy of David Jonker



# CONTENTS

Introduction xvii

## PART 1 Overview

### Chapter 1 Why Graphs? 3

Visualization in Business 4  
Graphs in Business 7  
    *Finding Anomalies* 9  
    *Managing Networks and Supply Chains* 11  
    *Identifying Risk Patterns* 15  
    *Optimizing Asset Mix* 18  
    *Mapping Social Hierarchies* 20  
    *Detecting Communities* 22  
Graphs Today 25  
Summary 26

### Chapter 2 A Graph for Every Problem 27

Relationships 28  
Hierarchies 32  
Communities 36  
Flows 40  
Spatial Networks 45  
Summary 49

## PART 2 Process and Tools

Process 52  
Tools 53

### Chapter 3 Data—Collect, Clean, and Connect 55

Know the Objective 56  
Collect: Identify Data 56  
    *Potential Graph Data Sources* 57

*Potential Hierarchy Data Sources* 65

*Getting the Data* 67

Clean: Fix the Data 69

Connect: Organize Graph Data 71

*Compute the Graph* 73

*Graph Data File Formats* 75

Putting It All Together 85

Summary 85

### Chapter 4 Stats and Layout 87

Basic Graph Statistics 88

*Size (Number of Nodes and Number of Edges)* 88

*Density* 88

*Number of Components* 89

*Degree and Paths* 90

*Centrality* 93

*Viral Marketing Example* 95

Layouts 97

*Node-and-Link Layouts* 97

*Other Layouts* 98

*Force-Directed Layout* 99

*Node-Only Layout* 106

*Time Oriented* 107

*Top-Down and Other Orthogonal Hierarchies* 109

*Radial Hierarchy* 111

*Geographic Layout and Maps* 112

*Chord Diagrams* 114

*Adjacency Matrix* 115

*Treemap* 117

*Hierarchical Pie Chart* 118

*Parallel Coordinates* 118

Putting It All Together 122

Summary 123

## Chapter 5 Visual Attributes 125

- Essential Visual Attributes 127
- Key Node Attributes 129
  - Node Size* 129
  - Node Color* 132
  - Labels* 137
- Key Edge Attributes 143
  - Edge Weight* 143
  - Edge Color* 144
  - Edge Type* 144
- Combining Basic Attributes 146
- Bundles, Shapes, Images, and More 148
  - Bundled Edges* 148
  - Shape* 148
  - Node Image* 149
  - Node Border* 150
  - More Attributes* 151
  - Interference and Separation* 152
- Putting It All Together 153
- Summary 155

## Chapter 6 Explore and Explain 157

- Explore, Explain, and Export 158
- Essential Exploratory Interactions 160
  - Zoom and Pan (and Scale and Rotate...)* 162
  - Identify* 164
  - Filter* 166
  - Isolate and Redo Layout* 168
- More Interactive Exploration 171
  - Identifying Neighbors* 171
  - Paths* 173
  - Deleting* 174
  - Grouping* 176
  - Iterative Analysis* 176
- Explain 177
  - Sequence of a Data Story* 178
  - Legends* 180
  - Annotations* 181

*Export Data Subsets, Graphs, and Images* 183

- Putting It All Together 185
- Summary 186

## Chapter 7 Point-and-Click Graph Tools 187

- Excel 188
  - Summarizing Links* 188
  - Extracting Nodes* 190
  - Adjacency Matrix Visualization in Excel* 190
- NodeXL 193
  - NodeXL Basics* 193
  - Social Network Features* 196
- Gephi 201
  - Gephi Basics* 201
  - Caveats* 205
- Cytoscape 208
  - Cytoscape Basics* 209
  - Importing Data into Cytoscape* 210
  - Visual Attributes* 212
  - Apps Menu* 218
- yEd 218
  - yEd Basics* 219
- Summary 222

## Chapter 8 Lightweight Programming 223

- Python 224
  - Getting Started* 224
  - Cleaning Data* 225
  - Extracting a Set of Nodes from a Link Data Set* 227
  - Transforming E-mail Data into a Graph* 233
  - Graph Databases* 241
- JavaScript and Graph Visualization 242
  - D3 Basics* 242
  - D3 and Graphs* 250
  - D3 Springy Graph* 264
- Summary 272



## PART 3 Visual Analysis of Graphs

### Chapter 9 Relationships 275

- Links and Relationships 276
  - Similarities in Fraud Claims* 277
  - Cyber Security* 279
- E-mail Relationships 282
  - Spatial Separation* 283
- Actors and Movies 286
- Links Turned into Nodes 290
- Summary 292

### Chapter 10 Hierarchies 293

- Organizational Charts 293
- Trees and Graphs 297
- Drawing a Hierarchy 300
- Decision Trees 306
- Website Trees and Effectiveness 309
- Summary 314

### Chapter 11 Communities 315

- What Defines a Community? 317
- Graph Clustering 318
  - A Social Network Case Study* 319
  - Social Media Using NodeXL and Gephi* 320
  - Layouts that Cluster* 323
  - Using Color to Characterize Clusters* 326
  - Community Detection* 328
  - Using Color to Distinguish Clusters* 330
  - Community Topic Analysis* 334
  - Community Sentiment* 338
- Cliques and Other Groups 342
  - Cliques in Social Media* 343
  - Community Groups with Convex Hulls* 345
- Summary 348

### Chapter 12 Flows 351

- Sankey Diagrams 352
- Constructing a Sankey Diagram 356
  - Create the Page Structure* 357
  - Process and Model the Data* 358
  - Visualize the Data* 358
  - Highlight Flow through a Node* 362
- Community Layouts with Flow 364
- Chord Diagrams 367
- Constructing a Chord Diagram 369
  - Prepare the Data* 370
  - Create the Page Structure* 371
  - Process and Model the Data* 372
  - Visualize the Data* 376
  - Interactive Details on Demand* 382
- Behavioral Factor Tree 384
- Summary 387

### Chapter 13 Spatial Networks 389

- Schematic Layout 390
  - A Modern Application* 393
- Small World Grouping 397
- Link Rose Summaries 398
  - Building a Link Rose Diagram* 401
- Route Patterns 408
  - Visualizing Route Segments* 410
  - Track Aggregation* 414
- Summary 415

## PART 4 Advanced Techniques

### Chapter 14 Big Data 419

- Graph Databases 421
  - A Product Marketing Example* 422
  - Creating and Populating a Graph Database* 424
- Graph Query Languages 427
  - Gremlin for Graph Queries* 428

<i>Using Graph Queries to Extract Neighborhoods</i>	432	<i>Spatial Transaction Analysis</i>	469
Analyzing Neighborhoods	435	Summary	472
<i>Topic Word Clouds</i>	441	<b>Chapter 16 Design</b>	<b>473</b>
Plotting Network Activity	444	Nodes	474
Community Visualization	446	<i>Node Shape</i>	475
Summary	448	<i>Node Size</i>	484
<b>Chapter 15 Dynamic Graphs</b>	<b>449</b>	<i>Node Labels</i>	485
Graph Changes	450	Links	486
<i>Organic Animation</i>	450	<i>Link Shape</i>	486
<i>Full Time Span Layout</i>	454	Color	492
<i>Ghosting</i>	455	<i>Color Palettes</i>	492
<i>Fading</i>	457	Summary	496
<i>Community Evolution</i>	458	<b>Glossary</b>	<b>497</b>
Transaction Graphs	461	<b>Index</b>	<b>501</b>
<i>Clustered Transaction Analysis</i>	461		

# INTRODUCTION

This book is about the application of graph visualization and analysis for business. Graph applications are a unique and valuable resource for discovering actionable insights in data. In recent years, analysts inside some of the world's most innovative companies have been intensively exploring graph-based approaches to gain deeper understanding of the dynamics of their businesses while discovering opportunities and strategies for improvement.

As the volume, variety, and velocity of available data has grown, so has the need for techniques and technology to make sense of it all. Organizations have become acutely aware of the limitations of simple dashboard-style charts. Dashboards are good at showing metrics and trends. They can inform you when areas of business are underperforming or outperforming others, but they cannot begin to tell you *why*, and understanding why is key to taking effective action.

The function of a graph is to represent links between things, revealing the structure and nature of relationships in data. Relationships are fundamental to the why and the how of things, which is one of the reasons graph analysis and visualization has so much potential for value.

Looking back on 20 years of our personal history designing and building new applications for business and intelligence analysts, the authors realize that graphs have played a role in many of those solutions. Today, several of our most significant research and software development efforts are, in essence, graph-based.

Despite the utility of graphs, however, little has been published about the application of graphs outside of the world of science, and even less has been published about graph design. With recent advancements in the capabilities of open source graph tools and libraries, graphs have become accessible to every business analyst, but access to knowledge of effective principles and techniques for graph analysis and visualization remains relatively limited. Our hope in writing this book is to help change that.

## WHO THIS BOOK IS FOR

This book is for data scientists and analysts interested in applying graph analysis to decision-oriented problems. The examples provided are taken from the business world, but the principles and techniques used are highly relevant to government and non-profit problems as well.

No prior knowledge of graph theory or practice is required. A reader who is new to graph analysis should find it useful to read this book from start to finish. More experienced readers may choose to skip ahead to subjects of interest in Part 3, which expands in detail on specific analytic themes.

Some examples in this book include light programming, but the majority of sample applications use point-and-click tools. In both cases, a moderate level of technical aptitude will be required.

## HOW THIS BOOK IS STRUCTURED

This book is composed of four parts. The first part represents a broad introduction to the subject of graphs. Subsequent parts are organized into progressively more specialized or advanced topics. Chapters 3 through 10 are written by Richard Brath, and the remaining chapters by David Jonker.

- **Part 1**—In the first part of the book, the authors provide an overview of graph applications in business and introduce various types of graphs, which are covered in more detail in Part 3.
- **Part 2**—The second part provides a comprehensive look at the major steps in the process of graph visualization and analysis.
- **Part 3**—The third part of this book is organized into distinct analytic themes and associated graph types and techniques.
- **Part 4**—The fourth part focuses on advanced topics representing areas of ongoing research, as well as fundamental design principles.

## MATERIALS FOR DOWNLOAD

This book includes online data files, source code distributions, and graph visualization files to accompany the examples provided. These Supplemental Materials are organized by chapter. The software required to view or run these files is described in each of the chapter examples. Files for download include the following:

- **Data files**—Most data files are available in a generic format such as text (.txt) or comma-separated values (.csv), which can be read directly into graph software or otherwise used by programs. In some cases, there will be two files, one for nodes and one for edges (that is, the links between nodes). In other cases, graph data files will be provided in a graph-specific file format, such as .gdf or .graphml. These are formats that many graph tools import directly.
- **Excel files**—There are a few Excel spreadsheet examples identified by .xls or .xlsx file extensions. These require Microsoft Excel in order to run.
- **Graph visualization files**—Some examples also include graph visualization files such as .gephi or .cys. These are files associated with specific graph visualization software such as Gephi or Cytoscape, respectively. To view these files, you must first download the free graph visualization software package and install it. See the following section for details.
- **Python code**—Programming examples use the Python language. These programming files are identified by the extension .py. Python examples are done in version Python 3.x and require the download and installation of Python. See the following section for details.
- **HTML and JavaScript**—Examples using JavaScript are typically web pages containing JavaScript and identified as .html files. These files will run in a standard modern web browser such as the latest version of Chrome or Firefox.

Source code for the samples is available for download from the following website:

[www.wiley.com/go/GraphAnalysisVisualization](http://www.wiley.com/go/GraphAnalysisVisualization)

## WHAT YOU NEED TO TRY THE EXAMPLES

A variety of tools are used in the book to process data and/or visualize data. In order to use the data files previously identified, the following software may be required:

- **Gephi**—The end-user point-and-click free software product Gephi (<https://gephi.github.io/>) is used for many of the graph visualization examples in the book. Many of the data files can be imported into Gephi for analysis and visualization. Chapter 7 of the book discusses some of Gephi’s features, building on the basic graph analysis process described in Chapters 3 through 6.
- **Cytoscape**—Cytoscape ([www.cytoscape.org/index.html](http://www.cytoscape.org/index.html)) is another free end-user software tool for graph analysis used in many examples in the book. Many of the data files can also be imported in Cytoscape for analysis and visualization. Chapter 7 discusses some of Cytoscape’s features and also outlines some of the differences between Gephi and Cytoscape.
- **yEd**—yEd ([www.yworks.com/en/products/yfiles/yed/](http://www.yworks.com/en/products/yfiles/yed/)) is an alternative free end-user point-and-click software product made by yWorks for graph analysis and visualization.
- **Excel**—Microsoft Excel (<http://products.office.com/en-us/excel>) spreadsheets are used in several examples. Excel is not free, but most readers will already have a copy, and Microsoft does allow download for time-limited evaluations. Several examples also use the NodeXL plug-in for Excel.
- **NodeXL**—Excel allows developers to create plug-ins that access and enhance Excel’s functionality. NodeXL (<http://nodexl.codeplex.com/>) provides graph functionality for social network data retrieval, as well as graph analysis and visualization.
- **Python**—For programmatic manipulation of data, the Python 3 (<https://www.python.org/>) programming language is used in some examples. Python is freely available.
- **A modern browser**—While any modern web browser should be capable of viewing the JavaScript/HTML examples, Chrome ([https://www.google.com/intl/en\\_us/chrome/browser/](https://www.google.com/intl/en_us/chrome/browser/)) was the browser used by the authors.

- **D3.js**—D3 (<http://d3js.org/>) is a JavaScript library used to create a variety of interactive data visualizations in a browser, and used, for example, in Chapter 8.
- **Aperture JS**—Aperture JS (<http://aperturejs.com/>) is a JavaScript framework library used in some of the examples in the later part of the book, for example, in Chapter 12.
- **Titan**—A Titan (<http://thinkaurelius.github.io/titan/>) graph database is used for several big data examples found in Chapter 14.

To use these software libraries and tools, you will need to download them yourself and install them, with the exception of the JavaScript libraries, D3.js, and Aperture JS. These are packaged with the examples for download from the companion website specified earlier.

## CAVEATS

The chapters in this book use case study examples to illustrate various applications and forms of graphs and how to use them yourself. Illustrations make use of real tools and real data where possible. There are caveats to keep in mind with both of these.

While the authors have used open source tools that are freely available to anyone, many of these tools are still works in progress and, as such, lack some of the polish and robustness you might expect of a finished product. Expect that a little extra patience will, at times, be the price of being an early adopter. Another aspect of documenting work-in-progress tools is that they are more likely to change. Use the tool-related steps in this book as general guidelines to a process. If the user interface does not seem to be exactly as described, find the matching items in the newer interface. If you cannot find them yourself, a quick Internet search is usually enough to find what you're looking for.

The other caveat to keep in mind is about the data being analyzed. A book like this depends on public data sets. While immense strides have been made in recent years in opening up corporate data sets to the public for advancing the art and science of analytics and visualization, private data sets are invariably larger and richer. While the analysis in this book is true to the data used, in many cases the data is only a proxy or sample of what can be found inside a corporate network. Treat the analysis as a template approach that can be reproduced with access to all of your data.

## CONVENTIONS

To help you get the most from the text and keep track of what's happening, a number of conventions have been used throughout the book.

### WARNING

Warnings hold important, not-to-be-forgotten information that is directly relevant to the surrounding text.

### NOTE

Notes indicate notes, tips, hints, tricks, or and asides to the current discussion.

### TIP

Tips are hints or tricks to help you master the information being discussed.

As for styles in the text:

- New terms and important words are *highlighted* when introduced.
- Keyboard strokes are shown like this: Ctrl+A.
- Filenames, URLs, and code within the text are shown like so:  
`persistence.properties`.



# Graph Analysis *and* Visualization



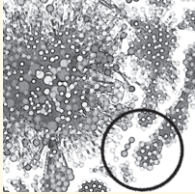
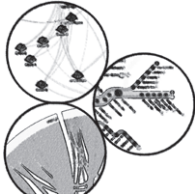
## Overview

The first part of this book introduces the subject of graphs and provides answers to two essential questions: why are graphs valuable to business analysis, and what kinds of opportunities can they be used to discover? A wide spectrum of techniques and applications are discussed, drawing from history and real-world experience. Case examples are used to illustrate value.

Before proceeding to a discussion of the process of graph analysis in the second part of the book, this overview provides you with a sense of just how many types of graphs there are and how many areas of potential value exist, even within a single business. References serve as a guide to subsequent chapters in the third part of the book, which cover each class of graph in more detail and step through tutorial style applications of graph analysis.

Table P1-1 describes the topics of Chapters 1 and 2.

**TABLE P1-1:** Overview

TOPIC	DESCRIPTION
<p>Why Graphs? (Chapter 1)</p> 	<p>What are graphs, and why are they useful to a business analyst? Chapter 1, “Why Graphs?,” introduces the concept of graphs, and defines several key terms used in this book. Select historical and modern anecdotes recount applications of graph analysis and visualization in business, documenting a steady rise to prominence spurred on by today’s challenges of vast and complex data. Real-world cases attest to the value of graphs.</p>
<p>A Graph for Every Problem (Chapter 2)</p> 	<p>Chapter 2, “A Graph for Every Problem,” provides a systematic overview of the wide variety of graph types and the kinds of problems they are useful for solving. The discussion begins with an example contrasting how relationships revealed in other ways can also be expressed using nodes and links. Subsequent topics describe graph techniques for gaining business insights involving hierarchies, communities, flows, and spatial networks. References are included to further detail in subsequent chapters.</p>

## WHY GRAPHS?

This book is about graphs and how graphs can be used to help solve business problems. When many people hear the word “graph,” they think bar charts or line charts, and rightly so, because those are also sometimes known as bar graphs or line graphs. This book is not about charts. This book is about the node-link diagram kind of graph.

At its essence, a *graph* is a structured representation of connected things and how they are related. As you will discover in the following chapters, graphs are capable of representing complex data in a way that an analyst can make sense of.

Because graphs have a long history in mathematics, discussions about graph analysis and visualization tend to include a lot of confusing esoteric terms such as *edge* and *degree*. This area of study responsible for this is generally known as *graph theory*.

For the discussions in this book, we use more universally accessible and less ambiguous terms where possible. For example, a *link* is a relationship between *nodes* and is typically drawn as a line. Nodes are entities (or essentially “things”) that are joined by links. Nodes are often represented visually by a circle.

An edge is another word for a link in graph theory, and the term *degree* becomes a little less opaque if you are familiar with the concept of *six degrees of separation*, popularized by the play and movie of the same name. But only a *little* less opaque, because not only can “degree” mean the minimum number of steps of separation between linked entities, it can also mean the number of link connections that a node has.

The glossary at the end of this book can serve as a cheat sheet if you find you need a little graph-theory-to-English translation.

In some circles, graphs are still viewed as abstract and difficult-to-understand constructs used mainly by scientists walking around with disheveled hair. Although graphs do have a long-standing tradition in scientific circles, the reality is that, when properly designed and executed, graphs can be one of the most intuitive ways to analyze information. There is a good chance you have used graph representations if you drew things in a notebook or on a whiteboard to think through or explain concepts—which is really a form of visualization.

More importantly, graphs provide a means of gaining highly unique and valuable insight from data. Graph analysis brings complex relationships to light, informing effective decision-making. Visualization is central to that process. Being able to see relationships visually is critical to understanding, whether they be characteristics of the raw data or specific features highlighted by graph analytics.

Information visualization exists for the sole purpose of understanding more, and in less time. Our brains are naturally wired to perceive and comprehend things visually. Reading is a time-consuming, sequential process, requiring the reader to mentally piece together an understanding. Pictures can convey information instantly, revealing complex patterns and outliers in easily digested ways.

There was a time when visualizations were drawn by hand after the painstaking gathering of data. But today, computer systems can harvest vast amounts of data and turn it into pictures in mere milliseconds, enabling analysts to instantly comprehend and act on information. Virtually any business can now benefit from visualization, and, as a result, it has become core to systems across all industries and around the world. Graphs, however, are one of the last forms of visualization to remain underutilized. There was a time, though, when that was true for all information visualization in business.

## VISUALIZATION IN BUSINESS

The use of computer-rendered visualization for decision-making in business is a relatively recent phenomenon. Twenty years ago, as recent grads from the University of Waterloo

School of Architecture, we decided to abandon the design of physical landscapes for the lure of an emerging and wide-open new world of virtual landscapes. One of us spent a few years working on three-dimensional (3-D) modeling software before we joined forces with other colleagues to see if similar technology could be applied to the problem of displaying large amounts of abstract information for high-flying decision-makers in finance and other industries. The seeds of that collaborative venture were to grow into an eventual long-term partnership, which included William Wright and another young architect, Thomas Kapler.

In the early days of this venture into business visualization, the value of even primitive charts was not always widely understood or accepted in offices of Fortune 500 companies. Our first pitches to corporate decision-makers started with the most basic of value propositions—that of the value of visualization itself. The pitch started with a slide presenting a small table of numbers and a challenge to the executives in the room to describe patterns. The next slide followed with the same numbers shown in a line chart. Visualized, patterns were immediately clear. In the table, the patterns were clearly not. That basic principle was the foundation for extrapolating how visualization could be even more essential in gaining insights from data that was orders of magnitude bigger and more complex.

At that time, the use of computers for primitive charting was still in its infancy, and beyond that, a product industry for analyzing business data visually was (by and large) yet to be born. What little advanced work that was going on was confined to a handful of corporate research labs and start-ups. Business was uncharted territory, in all senses of the word.

In those early days, one of the obstacles to the adoption of visualization in the business world was the limited graphic capabilities of computer systems at the time. When Edward Tufte's book *Envisioning Information* (Cheshire, CT: Graphics Press, 1990) was published, best-practice examples in the industry were still print-based, and the case studies in his seminal design book were no exception. The average computer was still far behind in quality of display.

When we hit the streets of New York in the early 1990s with novel interactive 3-D demos for financial analysts and traders, they had nearly a hundred pounds of specialized hardware in tow. Powering a single system required a hefty Silicon Graphics Inc. (SGI) computer and monitor. Between wrestling the equipment in and out of taxi trunks, and

careening it down city sidewalks on rickety, collapsible hand carts, it didn't take long before a new machine received its first patch of duct tape.

The bigger problem was that pretty much no one on Wall Street (or the rest of the business world, for that matter) had an SGI machine. Interactive visualization software systems were a hard sell when they came with a five-figure price tag per user for a new machine and operating system that didn't run any of their other apps. We generated a lot of buzz making one-off prototypes for a long list of high-profile firms, but progression to wide deployments were hard to come by.

When Microsoft Windows computers finally began to roll out with improved graphics application program interfaces (APIs) and graphics cards, it was a game changer. Access to higher-quality graphics capabilities on most desktops removed the requirement for expensive specialized machines, representing a major step in the democratization of advanced visualization for business use. By the mid to late 1990s, widely deployed high-powered analytics client platforms like the Bloomberg Terminal were running on PCs. Even highly specialized and demanding systems like the NASDAQ MarketSite broadcast wall were run on commodity Windows computers.

As the graphics capabilities of hardware began to mature, awareness of the value of visualization also matured. Timely, accurate, quickly perceived events and trends were critical to making lightning-fast decisions on the trading floor and elsewhere where systems and events needed constant monitoring. In business analysis as well, the value of representing information graphically to aid insight and to support strategic-level decision-making was quickly gaining momentum across all industries.

Surrounded by a rapidly growing market, we found our niche at the fresh and exciting edge of uncharted territory. For example, when the NASDAQ MarketSite began its move from the private confines of a downtown office to a public studio on Times Square, rebuilding its software infrastructure in the process, we were granted the task of designing and building the visualization systems and content. To open on the eve of the Millennium, the new studio would be composed of a 40-foot-long broadcast wall made up of roughly a hundred displays, and an electronic display wrapping the seven-story exterior tower. More than 6,000 stocks and indices would be displayed visually on demand in real time for reporters and the general public.



Before and since then, we have found ourselves with the privilege of working behind the scenes to help many of the world's most innovative companies and organizations solve their toughest information problems visually, through design and technology development. In doing so, we have had an opportunity to witness how the industry has evolved inside the walls of almost a hundred businesses, spanning the most data-intensive of industries. As time has progressed, the volume of available data has only increased, and so has the latent potential of information that can be gained from it. Data is now literally everywhere, waiting to be tapped for actionable insights.

As the realization that visualization is needed to make sense of it all has grown, so has the realization that visualization systems must be highly interactive. It is not sufficient simply to plot data and view it, just as it is not sufficient to simply compute an answer and present it. *Analysis* is an interactive process of rapid query, answer, and exploration, involving computational processes, visual display, and visual manipulation. In the early 2000s, dissatisfaction with the perception of visualization as simply an output channel led the research community to coin the term *visual analytics* to better represent and promote the interactive sense-making aspects of analysis.

Another awareness that has grown with the increasing size and complexity of information problems in business is that a basic palette of line, bar, and pie charts is rarely enough to express all of the valuable information available, and to leverage it for decision-making. Richer forms and combinations of forms are needed. Graphs, as it so happens, are one of the most valuable.

## GRAPHS IN BUSINESS

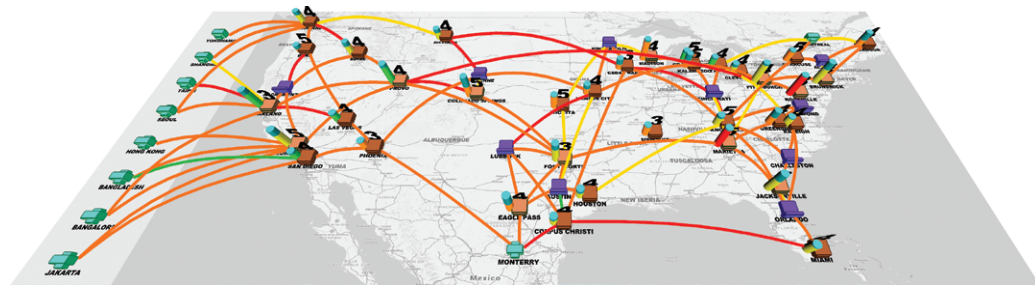
We have been helping organizations visualize and analyze graphs for almost 25 years. Graphs have been around much longer. One of the first graph problems was a deceptively simple question by Leonhard Euler: Was there a route so that each of the seven bridges in Königsberg, Prussia (now known as Kaliningrad, Russia), would be crossed only once, as shown on the left of Figure 1-1. Euler simplified the question into a graph, as shown on the right of Figure 1-1.

Since then, obviously many more problems have been analyzed as graphs, in business as well as science. Many such problems are geographic, just like Euler's.



**FIGURE 1-1:** In the seven bridges of Königsberg problem, Leonhard Euler explored whether each bridge could be crossed only once. On the left is a map showing the seven bridges, and on the right is the graph equivalent.

One of the first graph visualizations we produced was a geographic graph problem as well. In supply chain optimization, the task is to optimize the shipping between factories and warehouses to reduce costs. As shown in Figure 1-2, our visualization depicted the locations of facilities with icons indicating attributes such as type, inventory, capacity, and utilization, as well as major links indicating average costs.



**FIGURE 1-2:** One of the authors' first visualizations depicted a manufacturing and distribution supply chain network.

Various types of analyses can be done with this kind of supply chain visualization, ranging from inspecting individual routes to rationalizing the overall number of factories and warehouses. One interesting finding was that the costs between two particular factories doubled in March, June, September, and December. On inspection, it was discovered that a particular route was increasing shipping costs heavily at the end of each quarter. Further investigation showed that this route switched from land-based shipping to faster (but more expensive) air-freight shipping. Some questioning revealed that this change was driven by high-level objectives to reach quarterly targets. Because this pattern repeated consistently every quarter, the analysts realized that better planning and coordination between the two factories throughout the quarter could result in a better shipping schedule, and a reduction of shipping costs in the last month of the quarter. Similarly, graph analysis and visualization can be used in the analysis and optimization of other supply chain networks.

**NOTE**

Chapter 9, “Relationships,” discusses basic graphs and relationships in more detail.

## Finding Anomalies

*Spatial graphs* are often used to analyze the flow of goods around a company or around the world. One excellent early example of a flow graph is from Joseph Minard in the mid-1800s that, as shown in Figure 1-3, examined emigration around the world. Looking at it, you can easily see the flow of emigrants from the United Kingdom to the colonies, French and Germanic peoples to the United States, Portuguese to Brazil, as well as Africans, Indians, and Chinese to other locations.

Graphs can be made to analyze the movement of people, goods, or money, whether across the world, through processes, or through websites. Another of our early projects was for an airline company that wanted to analyze performance across its route network. Each link in the graph showed a flight route and had metrics such as revenue, passenger counts, efficiency, and profitability.