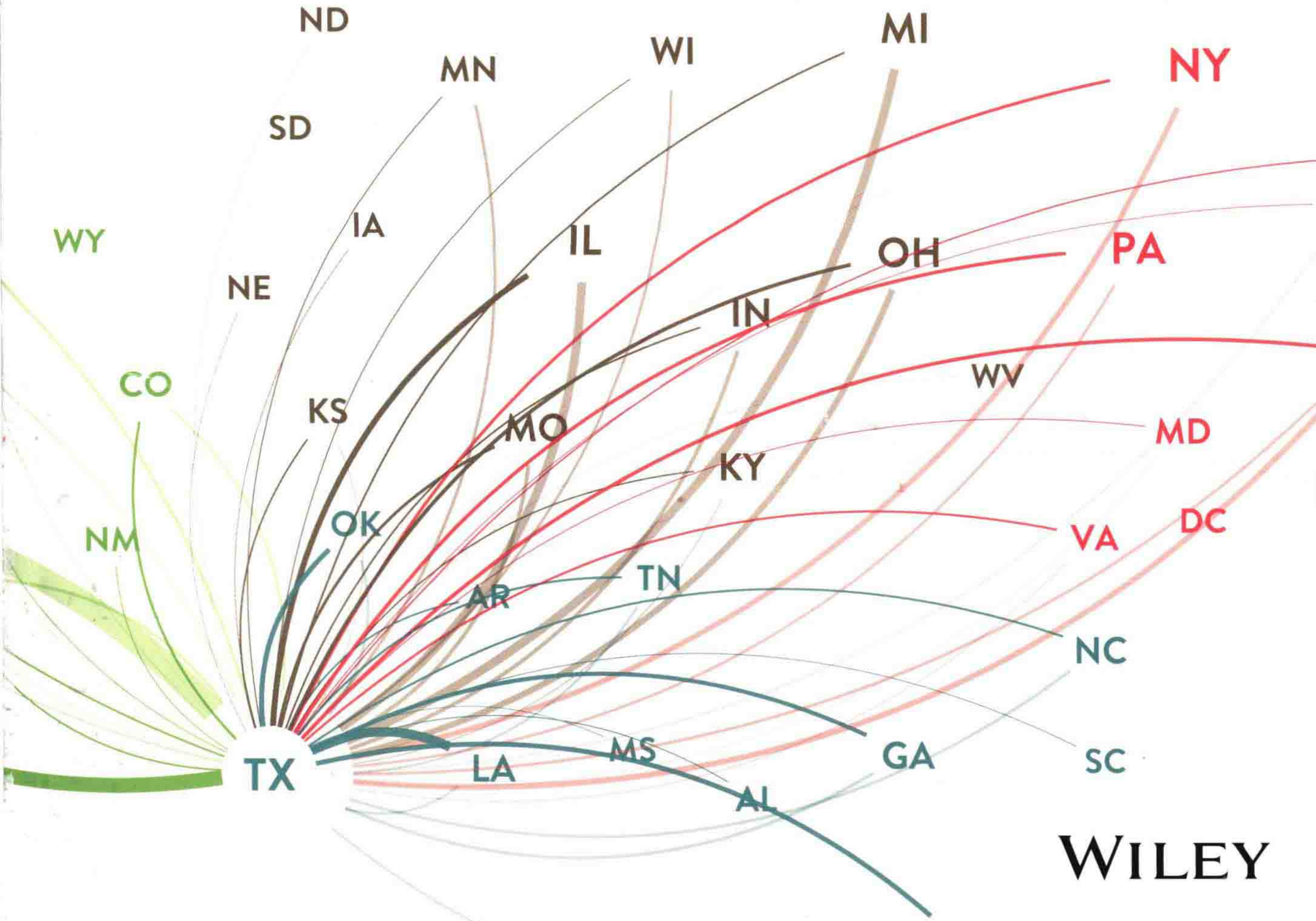
Graph Analysis and Visualization

Discovering Business Opportunity in Linked Data

Richard Brath and **David Jonker**



WILEY

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

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

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.

Library of Congress Control Number: 2014951021

Graph Analysis *and* Visualization

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 MacMurchy

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

Compositor
Maureen Forys,
Happenstance Type-O-Rama

Proofreader
Kim Wimpsett

Indexer
Johnna VanHoose

Cover Designer
Wiley

Cover Image
Courtesy of David Jonker

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 a 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`

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`) 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/`) 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. NodelXL (`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/`) was the browser used by the authors.