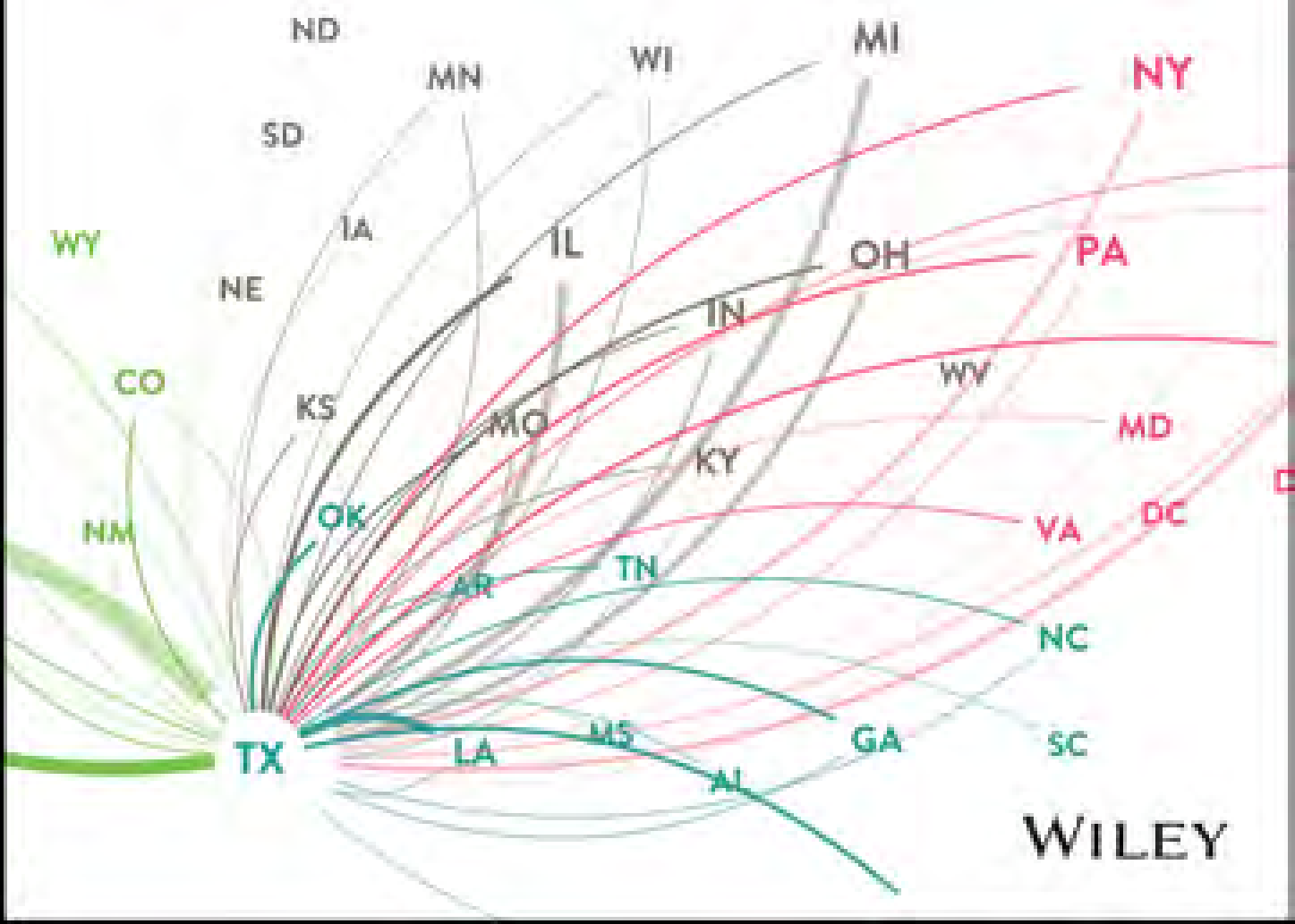
Graph Analysis and Visualization

Discovering Business Opportunity in Linked Data

Richard Brath and David Jonker



WILEY

CONTENTS

PART 1

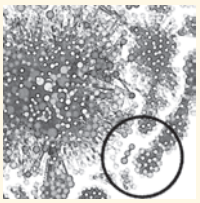
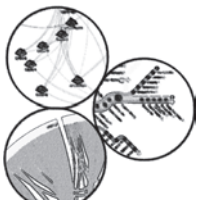
Overview

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
Why Graphs? (Chapter 1)		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.
A Graph for Every Problem (Chapter 2)		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.

1

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.

NOTE

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.

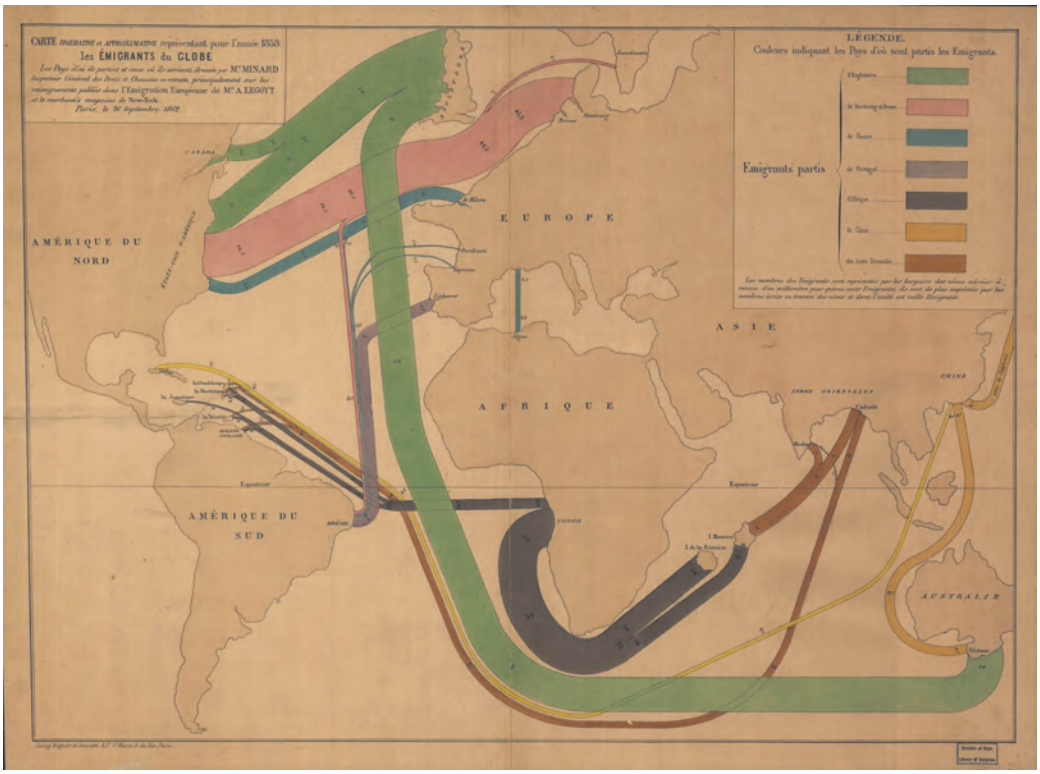


FIGURE 1-3: Joseph Minard's flow graph shows emigrants worldwide in 1858.

NOTE

A number of examples in this book look at statistics about movement between locations, specifically in the discussions in Chapter 12, "Flows."

Flow data sets, with an element of time, can quickly become Big Data. In such cases, we have used different strategies for dealing with these dynamic flow graphs, such as clustering. Figure 1-4 shows a recent application for investigating money flow between entities.

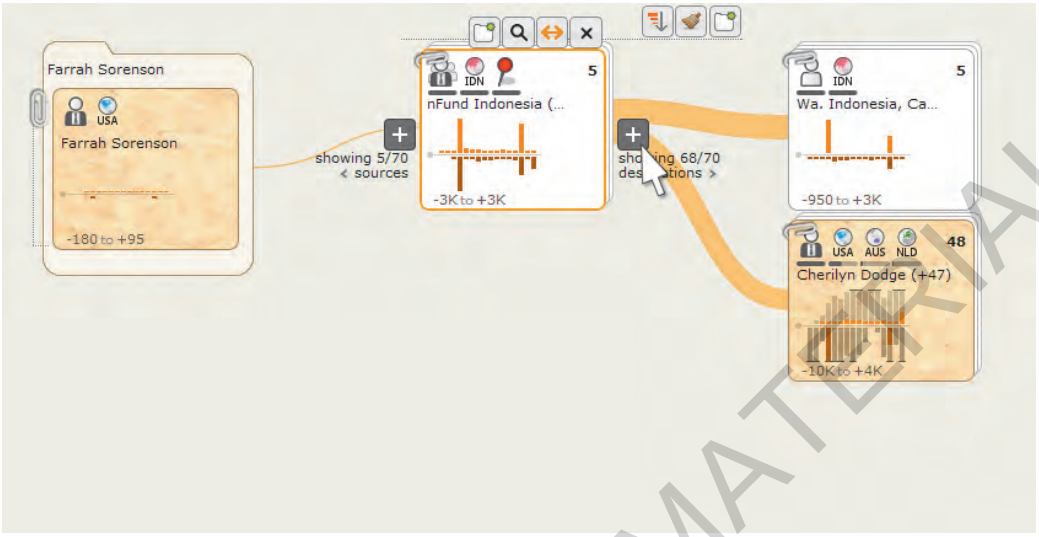


FIGURE 1-4: This flow graph shows money flow over time between different entities.

NOTE

This particular example is discussed in detail in Chapter 15, "Dynamic Graphs."

These graph examples are about finding and understanding anomalies such as unexpected links and unexpected flows. Identifying fraudulent activity and understanding paths through websites are examples of applications of this kind of graph analysis. Finding these anomalies can aid business by improving efficiencies, such as reducing losses or reducing clicks.

Managing Networks and Supply Chains

Pipelines, electrical systems, and railway networks are all large-scale physical networks. They are capital-intensive with large upfront costs that must be recovered through efficient operation. Similarly, large manufacturing and distribution networks have significant investments in plants, transport, warehousing, and other infrastructure. Adjustments must be made as conditions change.

Figure 1-5 shows an old diagram of freight traffic on a railroad from 1912–13. The thickness of various sections clearly indicates the volume of traffic, with two sides of each connection indicating the volume of traffic in either direction. If both sides are equal, then fully loaded box cars are generating revenue in each direction. Note the imbalance in freight traffic to and from Kansas City (top) and Ft. Scott shown here.

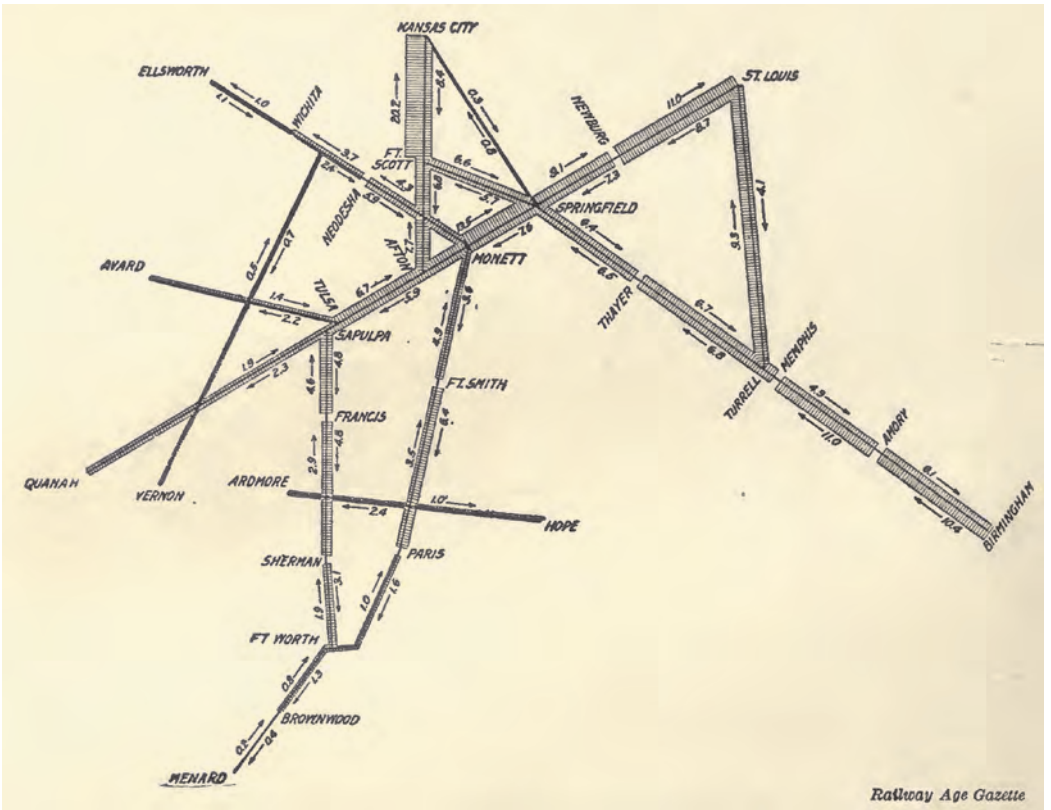


FIGURE 1-5: This graph shows freight traffic density and direction on the St. Louis and San Francisco Railroad in 1912–13.

*Image courtesy Prelinger Library (*www.prelingerlibrary.org*).*

Analyzing physical networks is an ongoing requirement for planners. As populations and energy use changes, the electrical grid must be adapted, too. Figure 1-6 shows a portion of the use of electricity on the West Coast of the United States from 2002. It shows only electrical transmission lines that are congested (that is, near capacity), potentially necessitating infrastructure upgrades.

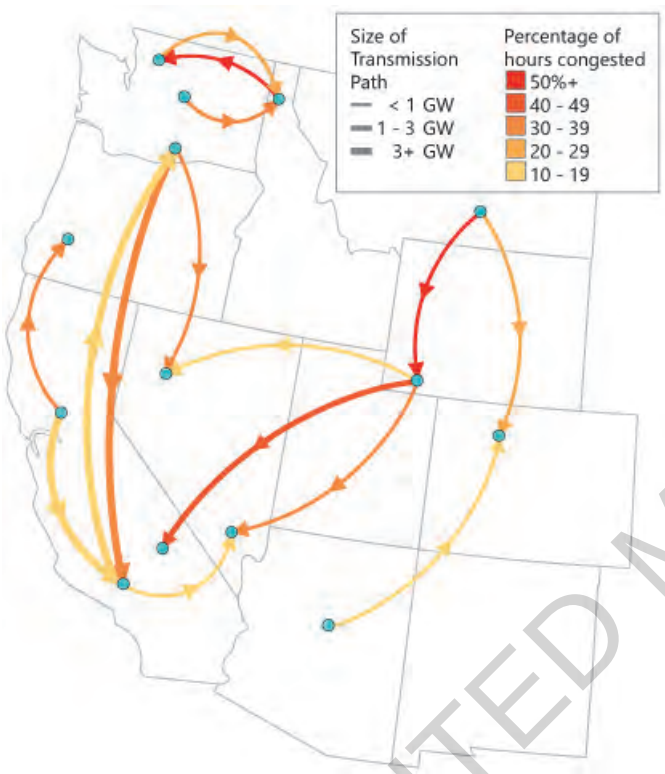


FIGURE 1-6: This shows congested transmission lines in the western United States electrical grid in 2002.

Redrawn based on U.S. Department of Energy National Transmission Grid Study 2002.

Many of these networks being analyzed in long-term planning must also be actively monitored to ensure efficient and trouble-free operation. One such project for us involved real-time data for a natural gas pipeline. In the case of the pipeline, nodes were compressor stations, and links were pipelines between each compressor station. Sensors in the compressor stations collected data such as pressure, flow, how close the compressor is operating to its limits, and alerts (such as a fault in a mechanical compressor). The alert-based system provided one way to easily monitor the system: no alerts equals no problem.

The solution provided was a graph visualization roughly along the lines of the one shown in Figure 1-7. The links were sized based on pipeline capacity, with nodes indicating flow through the station as a 3D bar, colored the node based on the limits (for

example, blue for not close to limit and red for close to, or exceeding, limit), and added a significant red flag on top of the node if an alert was triggered.

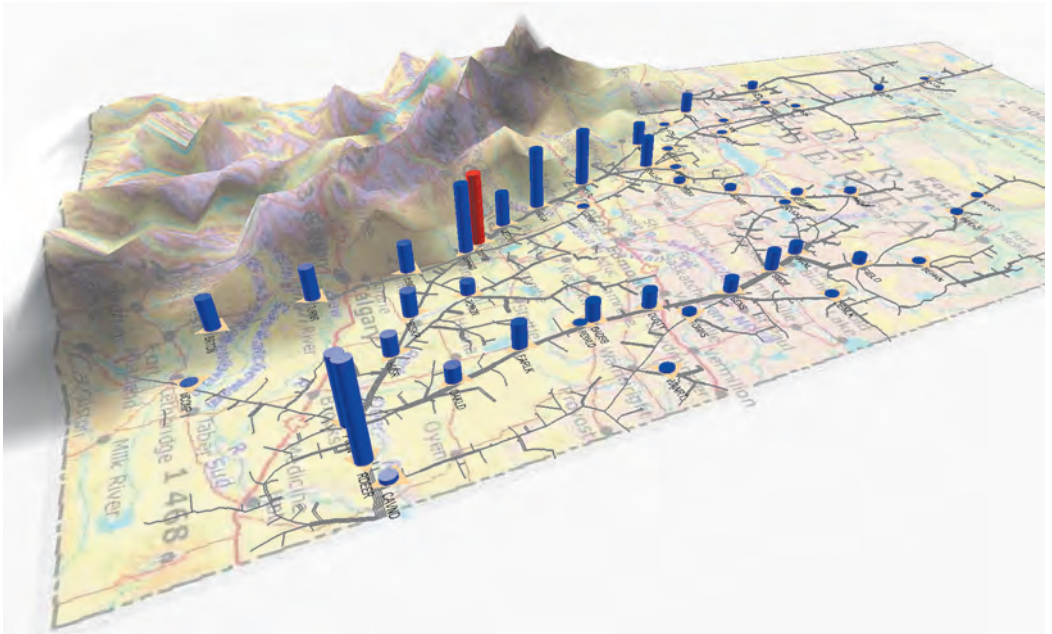


FIGURE 1-7: As shown in this pipeline graph, gas generally flows from the north (top of image) to the south (bottom left).

One interesting incident occurred shortly after we had completed an early version of the visualization. The alert system had no active alarms. But the visualization showed one compressor station operating close to it limits with a high volume going through the station (red in the figure). Inspecting all the node attributes associated with that station via a tooltip indicated no particular problems other than the station was working very hard and close to its limits. Having the whole graph visible meant that the analysts could visually inspect all the neighboring nodes for clues. One of the nodes immediately connected to this node had an extremely low volume (the low blue node immediately to the right).

The operators could easily see that the hard-working node was compensating for the neighbor node—in effect, performing additional work to maintain overall throughput of that portion of the network. This is a good example of where graph visualization is an effective complement to other kinds of graph analytics. The alert system by itself failed to

create a message for the problem node, but the visualization provided enough information that the viewer could see the problem and pinpoint its source.

> **NOTE** Geographic graphs are discussed more in this book, particularly in Chapter 13, "Spatial Networks."

Managing networks, regardless of real-time, daily, or monthly analysis, requires understanding multiple variables about both nodes and links in order to assess the overall network health. The graphic depiction of the network and the data acts as an aid to visually navigate hops to assess issues and understand their impact.

Identifying Risk Patterns

Beyond geographic networks, networks can simply be logical connections between things, such as computers or telephones. Figure 1-8 shows an early network drawing of the ARPANET (the forerunner to the Internet). One myth about the early ARPANET was that the network had many paths and decentralized message routing to deter nuclear attacks. However, this decentralization may have been more because of the unreliability of links and nodes in early computing.

Rather than focus on all the logical connections between specific computers, another way to look at the Internet is to examine where the traffic is going from and to. Particularly useful in network security is knowing which computers are targets for potential hackers and attackers, or otherwise performing actions in the network that are anomalous. This is a graph problem that can be drawn to show connections between the source computer (for example, the hacker or the internal thief's computer) and the target computer (for example, the corporate website or the offshore bank account).

Because many different kinds of events can occur (viruses, malware, bots, and so on), there are many different kinds of links. Furthermore, these network events are happening over time. They are transient, appearing and disappearing. There can be many different ways of representing this kind of graph, such as showing all the links, aggregating links by type of event, providing an interface to show links between only a set time period, and so on.

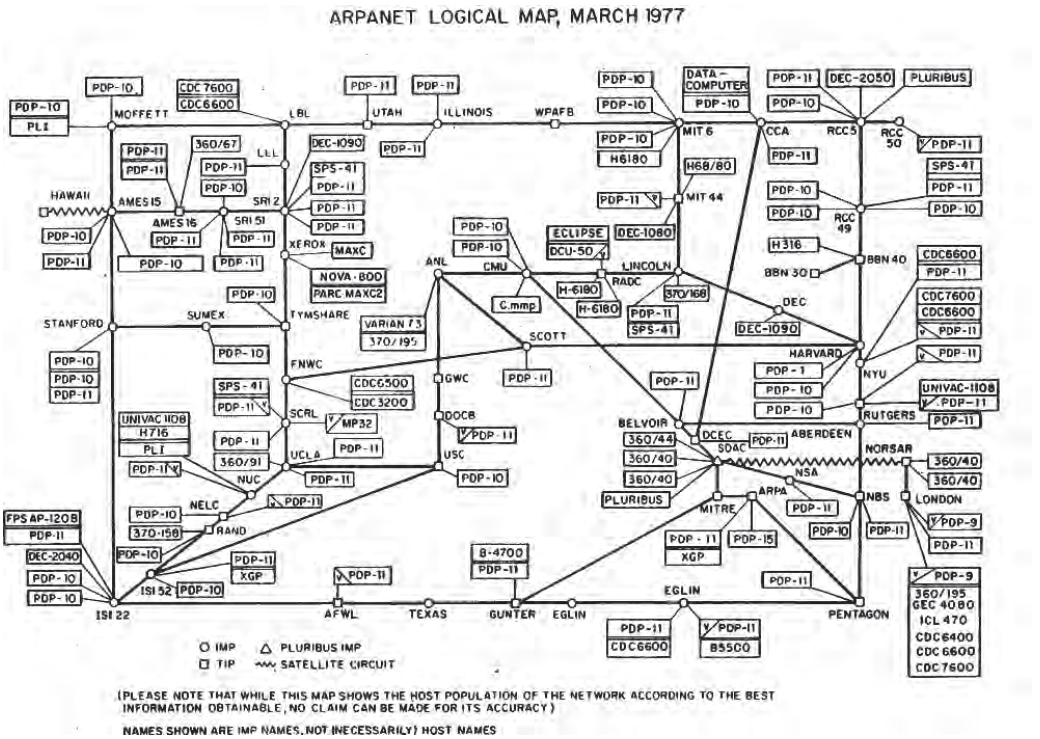


FIGURE 1-8: This early drawing from 1977 shows the ARPANET, the precursor to the Internet.

NOTE

Chapter 9 shows a few examples of graphs representing many links simultaneously.

Figure 1-9 shows a sketch of a graph visualization we constructed that used an interface to show only links over particular time periods. By isolating time periods, the viewer could identify event sequences, potential related events, and potential collaborators—symptomatic of a more organized attack. Also, different kinds of attacks have different visual signatures that stand out when viewing the patterns in a particular time period.

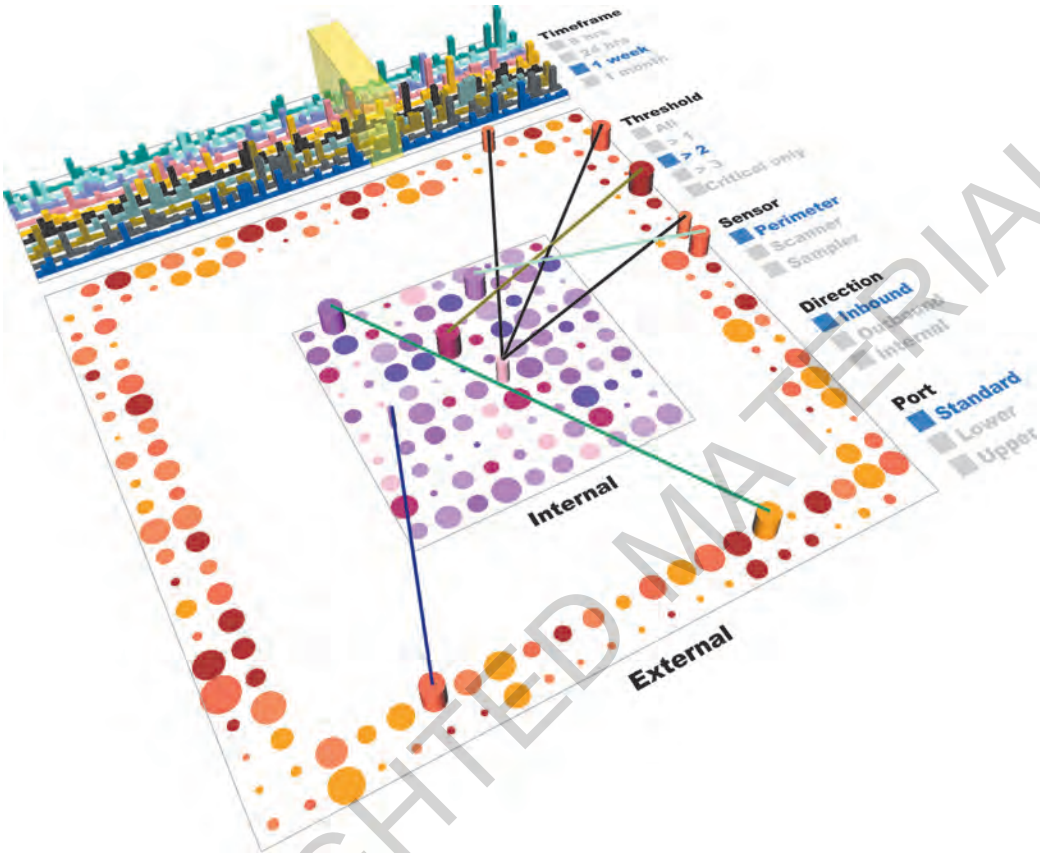


FIGURE 1-9: This graph visualization shows potential anomalies with connections between internal computers (inside) and external computers (around perimeter).

NOTE

Chapter 4, "Stats and Layout," discusses in detail the visual layout of a network.

Visualizing connections and patterns of connections may be useful for spotting risk, such as different types of threats to a physical network as shown here, as well as other types of risk, such as financial counterparty risk. Analyzing risk without graphs may lead to limited conclusions. Graph-based analysis can help reveal how risk exposure may extend to other entities.

Optimizing Asset Mix

The objective of a *market basket analysis* is to understand which products have a strong tendency to be purchased together. More generally, this is a graph where you are looking for strong correlations between things, which could be products purchased together, people who are popular at the same time, stock prices that move together, actors who appear in movies together, and so on.

One old approach to understanding these correlations was to create a matrix with each item listed in the columns and in the rows. The cells in the matrix indicate the strength of the relationship between the pair of items. When there are only a few items, the matrix can show all the possible connections between any pair of products, as shown in Figure 1-10.

Cross-Sell Patterns		First Device Purchased				
		Smart Phone	Music Device	Tablet Computer	Laptop Computer	Desktop Computer
Additional Device Purchased	Smart Phone	-	63%	12%	7%	28%
	Music Device	4%	-	1%	3%	2%
	Tablet Computer	19%	18%	-	18%	19%
	Laptop Computer	11%	6%	11%	-	3%
	Desktop Computer	4%	9%	8%	4%	-

FIGURE 1-10: This adjacency matrix shows how many times one product purchase leads to the purchase of the second product.

> **NOTE** Chapter 7, "Point-and-Click Graph Tools," discusses adjacency matrices in a bit more detail.

As the number of products grows, however, the number of potential connections is exponential. A matrix is less effective when looking at hundreds of items. To address that we have put together visualizations for problems which include analysis of market baskets of products at retail stores, the connections between people via e-mail, and the correlation of stocks.

In one fun example, we took a market basket visualization that we created for a client to compare correlations of financial assets and changed the data to a set of correlations

between some of the top Twitter celebrities, as shown in Figure 1-11. The distance between any pair of nodes indicates the strength of the correlation (close nodes have a strong relationship). Because there are many items, we turned off all the links to keep the display clean. Perhaps not surprisingly, there is a strong correlation between celebrities such as Justin Bieber, Lady Gaga, Felicia Day, and Taylor Swift. Inverse correlations are on the flip side in this visualization, and perhaps unsurprisingly, Margaret Atwood and Richard Florida are inversely correlated to the pop stars.

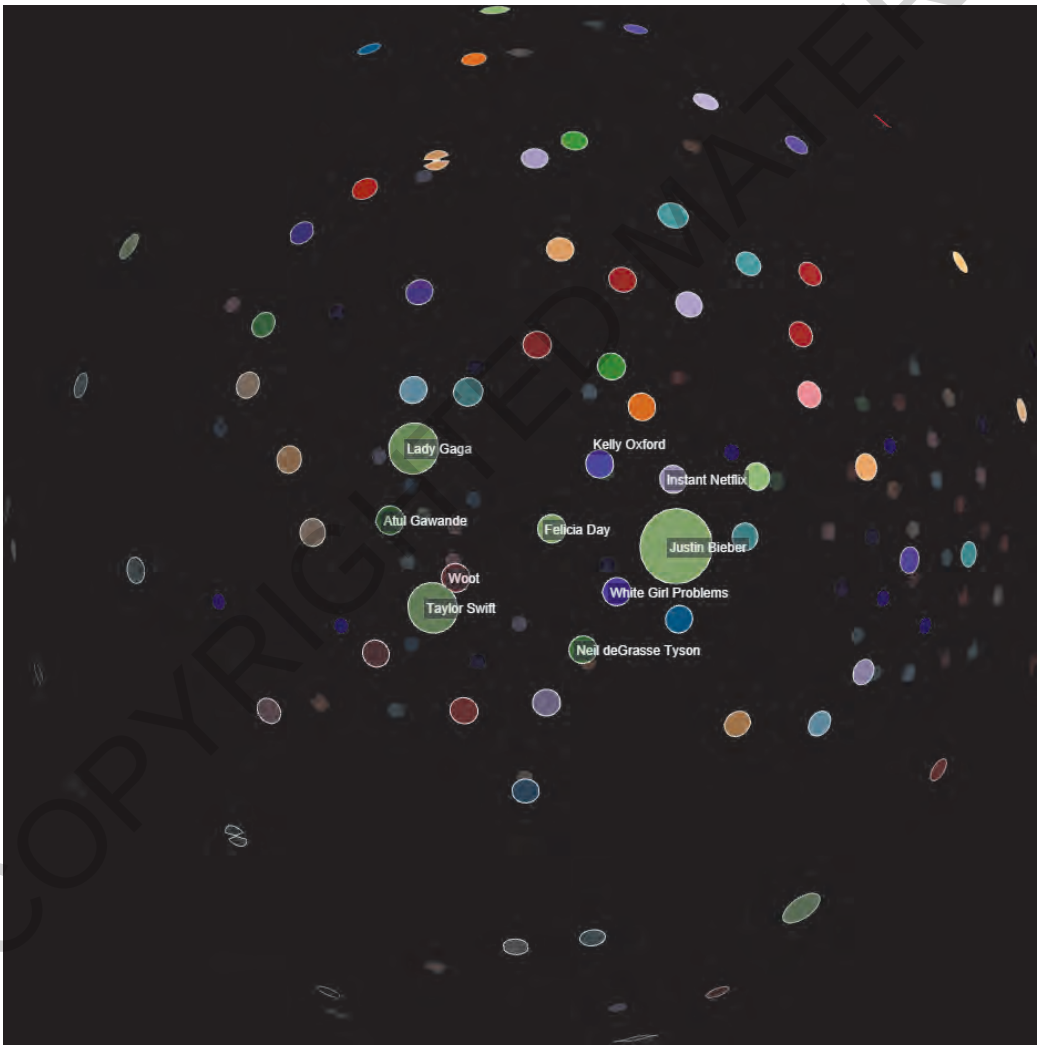


FIGURE 1-11: This shows that the correlations between top Twitter accounts (Justin Bieber, Felicia Day, Lady Gaga, and Taylor Swift) are all close together.

NOTE

Chapter 6, "Explore and Explain," discusses more about market basket analyses.

While an analysis of correlations between celebrities may seem trifling, a similar approach is used to optimize portfolios of other types of assets, such as financial portfolios, pharmaceutical drugs, or oil wells. The proximity of nodes as a result of force-directed layout algorithms (discussed in Chapter 4) provides insights into the asset choices that comprise of a collection of assets, such as close alternatives, isolated singletons, and opposites.

Mapping Social Hierarchies

There is a lot of current interest in social networks. Mapping out social networks goes back hundreds of years.

Figure 1-12 shows the genealogical tree for French royal family from Louis XIV to Louis XVI from the book *A Complete Genealogical, Historical, Chronological, And Geographical Atlas* by M. Lavoisne (Philadelphia: M. Carey and Son, 1820). This wonderful visualization shows direct rulers, spouses, offspring, and branches that merge together again. Nodes are people, with kings shown as crowns, men shown as filled circles, and women shown as transparent diamonds. Links are lines with time proceeding from top to bottom, and horizontal line style differentiates between the children of married spouses (plain line) or mistresses (diamond line).

NOTE

Chapter 5, "Visual Attributes," explores how to use visual attributes such as shape and color. Chapter 16, "Design," discusses related design considerations.

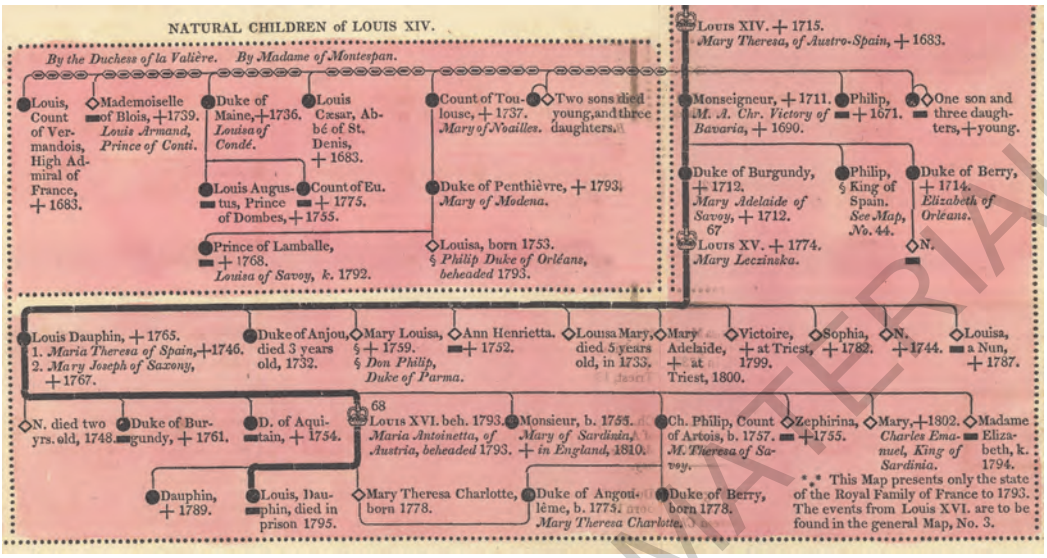


FIGURE 1-12: This portion of a genealogy chart shows the French royal family from Louis XIV to Louis XVI.

Courtesy davidrumsey.com.

In business environments, *organizational charts* (sometimes called *org charts*) are similar to genealogical trees. Although simple org charts work for small hierarchies, other approaches are needed for exploring large hierarchies with thousands of managers or tens of thousands of staff in contact centers. By combining the hierarchical view with time series views, trends and changes in performance can be viewed at any level as a time series, and up and down the hierarchy.

Figure 1-13 shows an early version of a visualization we created for a client showing org charts with time series. Consistent coloring across the nodes and links allows the viewer to track how the positive and negative contributions roll up.

> **NOTE** Chapter 10, "Hierarchies," provides more information about organizational charts.

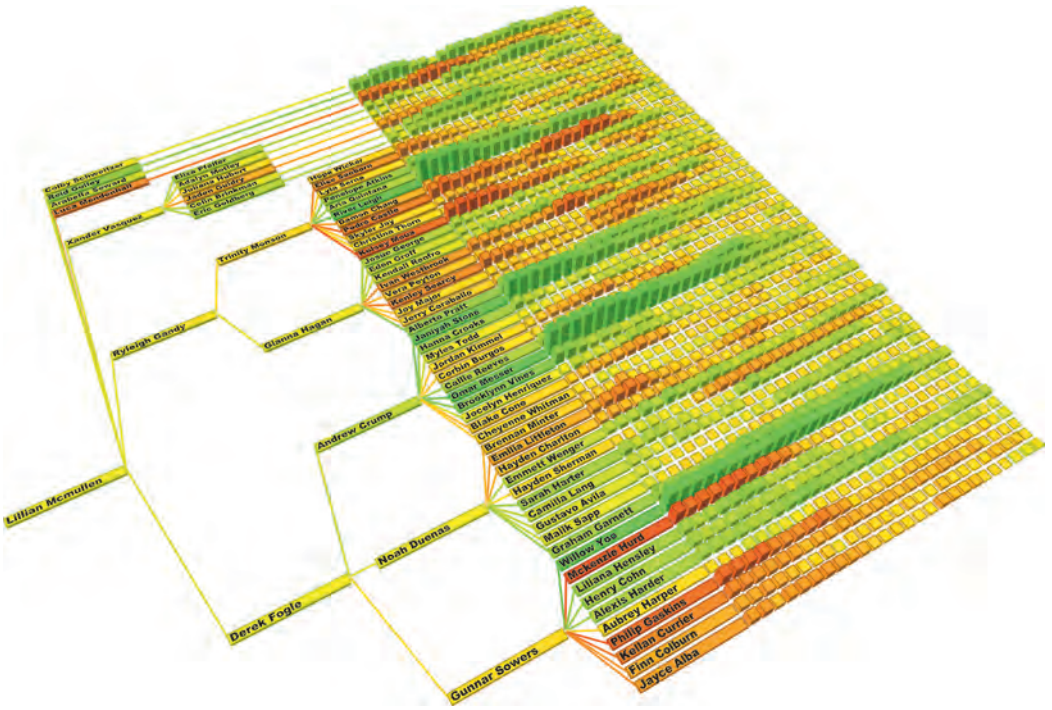


FIGURE 1-13: The left side of this organizational hierarchy uses color to indicate performance through all the levels, with the lowest level expanded on the right side to show performance over time.

Hierarchies are a unique type of graph and can be used to drill down through the organization to assess where the contribution to performance is coming from—for example, based on staff (as shown here) or based on other means such as attribution models. By providing this hierarchical decomposition, management can spot whether issues are localized, within a group or broad-based. Using this insight, they can respond more effectively to these different scenarios.

Detecting Communities

Beyond genealogical charts and the visualization of friend networks, visual analysis of social networks has many other applications. In health care, social networks can be used to analyze relationships and the potential spread of disease. Researchers have mapped out all the "romantic and sexual relationships" in a Midwestern high school (research paper: "Chains of affection: The structure of adolescent romantic and sexual networks" by Bearman, Moody, and Stovel). Out of 832 participating students, 573 were involved

in a sexual or romantic relationship. Of those, many (126) were involved with only one partner over the previous 18 months, but there were also larger components where a person may have been involved with more than one other person.

Figure 1-14 (created using Gephi) shows a large component of 288 students linked by sexual relationship. This graph is important because it indicates how approximately 50 percent of sexually involved students could be linked in the diffusion of sexually transmitted diseases (STDs).

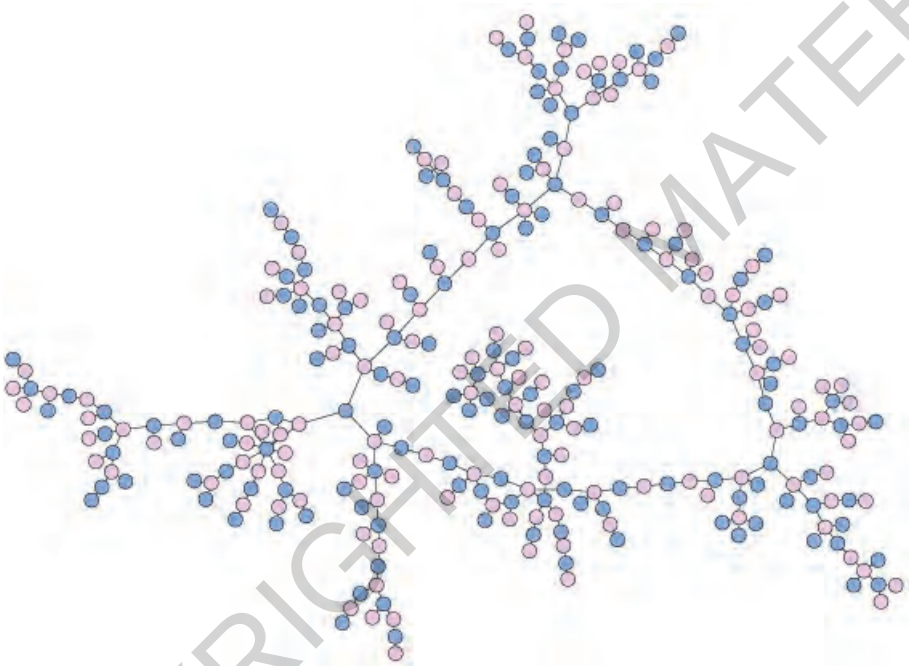


FIGURE 1-14: In this visualization of romantic and sexual relationships at a Midwest high school, you see how a large percentage of students surveyed are connected to each other through long chains of relationships.

The spread of diseases is similar to the spread of viral marketing or the spread of opinions and sentiment. Some firms may have data based on sales referrals or e-mail or extracted from social media such as Twitter.

Analyzing these social networks will often reveal clusters with higher densities of interconnections known in graph terms as *communities*. Identifying these groups of people and how they are connected can help a company identify different customer segments and better understand the dynamics of influence within and between them.

In a large company, social network problems may easily involve millions of nodes. Representing these graphs visually and exploring them for the purposes of extracting meaningful information is exceptionally difficult. Common desktop tools like Gephi (which are limited by in memory processing on a single machine) are not designed for graphs of that size.

We are involved in an ongoing advanced research effort exploring the use of cluster computing for community-detection and graph-drawing techniques to achieve highly scalable zoomable graphs with millions of nodes and tens of millions of links. Figure 1-15 shows an example of one such graph involving referrals. Clusters of medical practitioners seeing the same patients are outlined with circles, indicating communities.

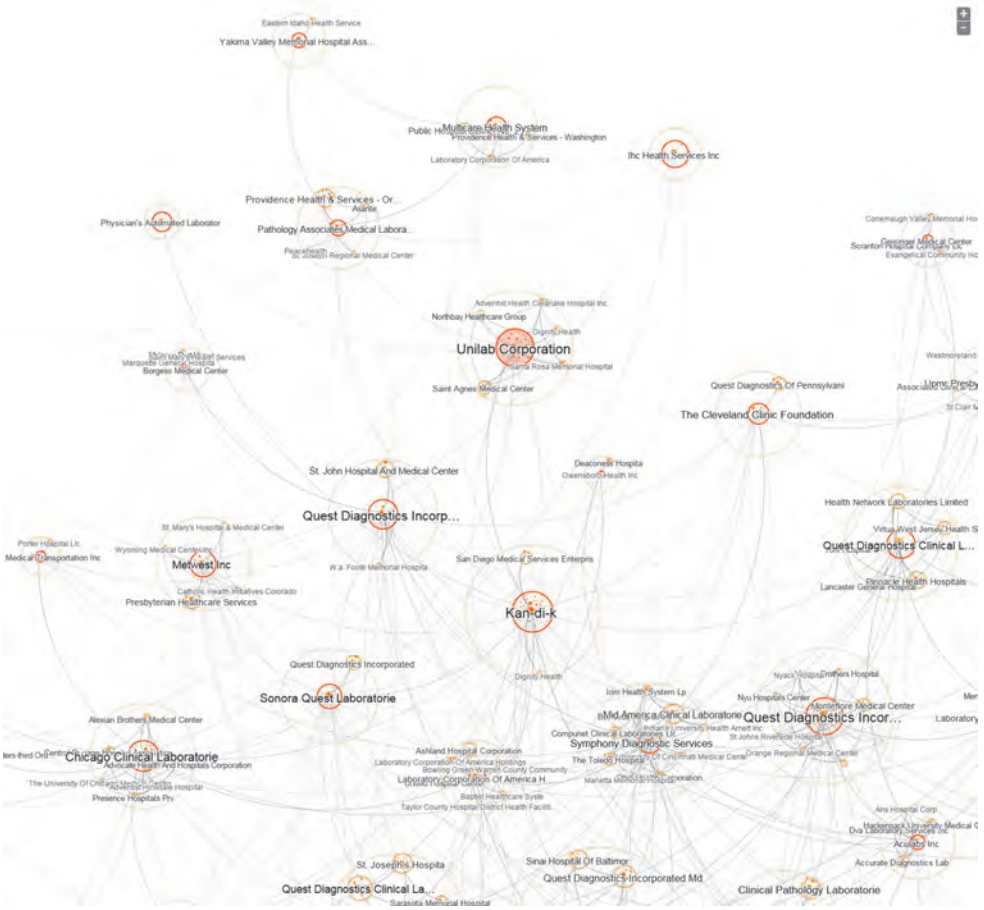


FIGURE 1-15: Use of distributed community-detection techniques and multi-scale graph drawing techniques can reveal community structure in very large graphs. Here, the DocGraph data set is visualized in its entirety, comprising millions of medical practitioner nodes and tens of millions of referral links.

NOTE

More information about the data used in this example can be found on the DocGraph project website at `http://docgraph.org`.

NOTE

The analysis of clusters and communities is outlined in Chapter 11, "Communities," and the theme is picked up again in Chapter 14, "Big Data."

Analysis of social networks can provide insights into clusters of people or organizations and influential connections within and between those clusters. These insights can be used to understand diffusion through a network (such as spread of coupons or a virus) and to understand communities (such as customer segmentation based on connections).

GRAPHS TODAY

In the age of Big Data, many of the world's most data-rich businesses are searching for new ways to make sense of vast streams of complex, irregular, sometimes unverifiable, interconnected data. Graph analysis and visualization is gaining momentum as a tool for helping to do just that. Graphs are particularly good at characterizing complex, compound relationships that are not easily described in black-and-white terms. They are also a natural choice for displaying networks, which are an increasingly integral part of many business data sets.

Desktop tools like Gephi and Cytoscape (which typically originate in scientific communities) have made strides in visual quality and scale for graph visualization and analysis. With their open and extensible nature, these tools can be easily applied to business problems, given the right amount of technical training and determination. With the prospect of cloud-based systems on the horizon, graphs promise to become even more easily accessible to the wider community of business analysts.

The goal of this book is to inspire creative thinking about the potential application of graphs to your own business problems and to share a little of our own domain knowledge in the hopes that you may try it yourself. Step-by-step tool usage and code samples are provided using case examples that demonstrate how graph analysis and visualization can be used to gain insights from data.

SUMMARY

Graph analysis is a powerful tool for discovering valuable information about relationships in complex data, representing significant business opportunity. Graph visualization is essential and, when used properly, can also be extremely intuitive. Information visualization takes advantage of natural perceptive abilities to allow an analyst to see more information, more quickly.

The importance of visualization in business has risen to widespread recognition as the volume of data available continues to increase. During that time, graphs have developed into an instrumental tool, with applicability in areas such as network monitoring, market basket analysis, influence analysis, and optimizing of processes and organizational structures. With the rise of Big Data the importance of techniques suited to dealing with complex relationships has risen with it. Need has fueled technology development, and today graph tools are emerging as a valuable resource available to any business analyst.

Chapter 2 provides a detailed overview of the many kinds of graphs and how they can be used in solving various business problems. The first example provides an illustration of how graphs are effective at intuitively summarizing relationships at a high level, while providing additional levels of detail with further analysis. Additional examples show different forms of graphs, as well as their relative strengths and suitability to answering specific kinds of questions.

GRAPH ANALYSIS AND VISUALIZATION:

DISCOVERING BUSINESS OPPORTUNITY IN LINKED DATA

Richard Brath & David Jonker

9781118845844

March 2015

$50.00 / £33.99

www.wiley.com/buy/9781118845844