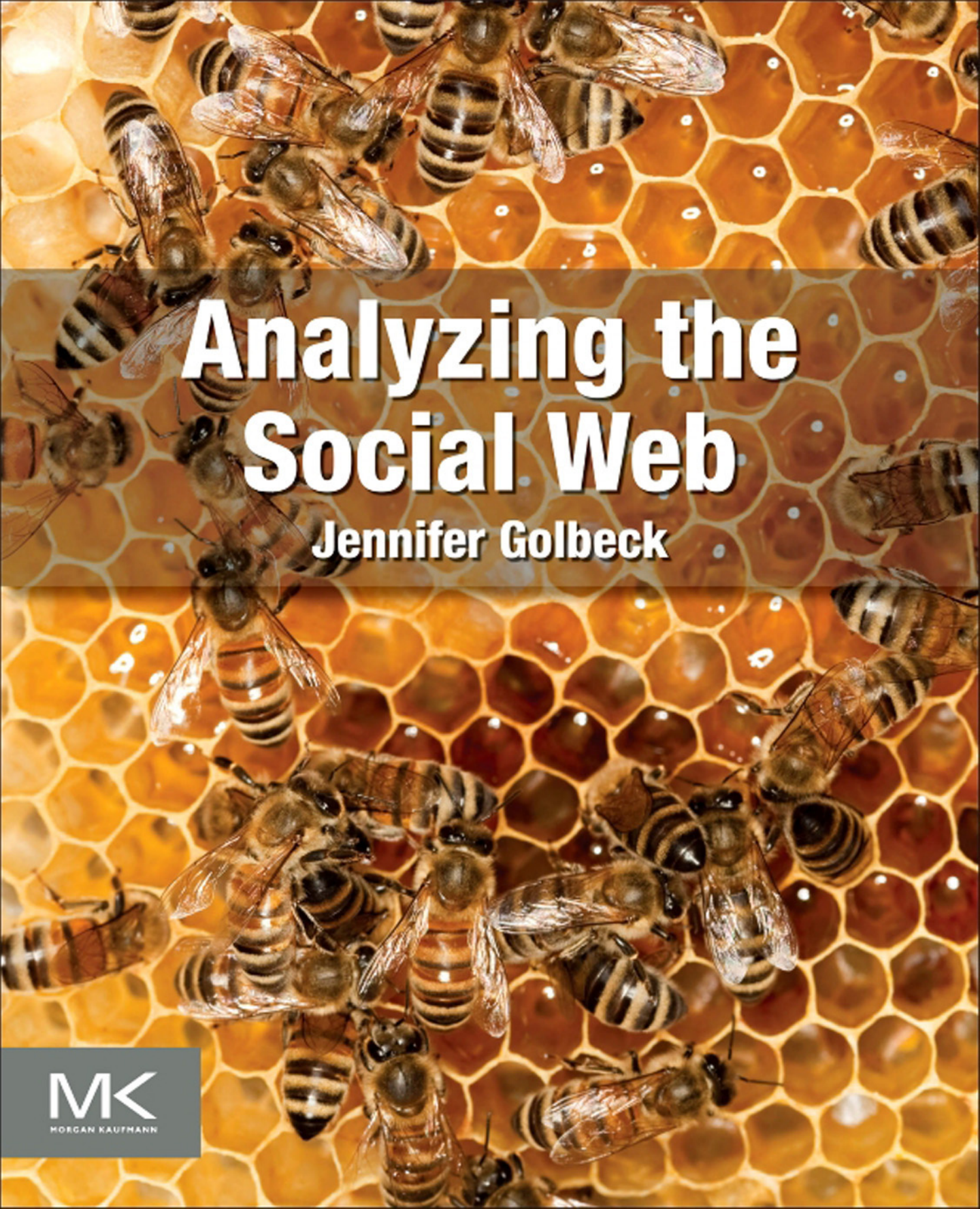

Analyzing the Social Web
Jennifer Golbeck
MK
MORGAN KAUFMANN

Analyzing the Social Web

This page intentionally left blank

Analyzing the Social Web

Jennifer Golbeck

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Morgan Kaufmann is an imprint of Elsevier

Acquiring Editor: Steve Elliot
Editorial Project Manager: Lindsay Lawrence
Project Manager: Punithavathy Govindaradjane
Designer: Mark Rogers

Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods or professional practices, may become necessary. Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information or methods described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
Golbeck, Jennifer.
Analyzing the social web/Jennifer Golbeck.—First edition.
pages cm.
Includes bibliographical references and index.
ISBN 978-0-12-405531-5 (alk. paper)
1.Social media. 2.Social networks. 3.Human-computer interaction. I. Title.
HM742.G65 2013
302.3—dc23
2012049046

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-405531-5

Printed and bound in the United States of America

13 14 15 16 17 10 9 8 7 6 5 4 3 2 1

For information on all MK publications visit our website at *www.mkp.com*

To Ingo

This page intentionally left blank

Brief Table of Contents

Table of Contents

This page intentionally left blank

List of Figures

Acknowledgments

Thanks to my students who served as guinea pigs, using drafts of this book in class, especially Sophal Chhay, Pano Papadatos, Esther Hwang, Daniel Osborne, Rasmeyleina Samel, Sarah Webster, Tansy Peplau, and Raj Zachariah. They all offered comments, corrections, and suggestions on the early drafts.

I am especially grateful to two people for their help with this text. Derek Hansen, a colleague whose presence at UMD I sorely miss, was a technical reviewer for the book. His comments were invaluable and guided this text to become something much better than it would have been without him. Tony Rogers also read every chapter as I wrote it, and edited it into a much more coherent and eloquent text. I am grateful to you both.

Noshir Contractor and Jim Hendler offered their guidance and insights into the development of the text. I'm honored to have had the help of two such prominent and well-respected researchers.

Thanks to other friends and colleagues for their input and advice, including Carman Neustaedter, Bryan Dennis, Marc Smith, Ben Shneiderman, Jenny Preece, Allison Druin, Ben Bederson, and all the members of the Human-Computer Interaction Lab at the University of Maryland.

And to Pi, K, and Ingo who supported me throughout this process.

This page intentionally left blank

Foreword

One way to track civilization's progress is by our dramatically increasing capacity to perceive, understand, measure, and predict the influences on our lives.

When mapmaker Gerardus Mercator (1512−1594) developed a rectilinear projection of parallels of latitude and meridians of longitude, he enabled travelers to perceive a comprehensive world view. Mercator's maps revealed the relationships among countries, giving travelers the capacity to measure distances and the tools to predict future positions for ships that maintained a fixed heading.

Another visual breakthrough was Renee Descartes's (1596−1650) infinite plane with x and y positions for attribute pairs. These Cartesian coordinates permitted analysts to plot algebraic equations to make discoveries about slopes, intersections, and correlations. Modern applications are pervasive, including charts that enable physicians to plot a child's height and weight, so as to see growth patterns, clusters of similar children, as well as simple errors and meaningful exceptions.

When Isaac Newton (1643−1727) described gravity as the attraction between two masses, he could then develop formulas to quantify the forces and shift from primitive ideas of planetary motion to more modern explanations that had predictive power.

Similarly, Charles Darwin's (1809−1882) understanding of natural selection was a revolution that recognized the complex relationships among animal species, plant life, climate, and environmental forces. He drew a tree of life that showed ever richer differentiation among and contrasts across species.

These stunning conceptual breakthroughs are just a few of the well-known transformative innovations that also include subject categories for books, the periodic table of elements, or choropleth maps to show regional economic or health data.

Within the past century a major shift is the growing recognition that networks effectively represent organizational structures, communications patterns, publication citations, and environmental interrelationships. While a spider's orderly web is a visible network, the harder to see network of food webs is vital for understanding, measuring, and predicting how changes in food chains trigger environmentally favorable or destructive forces.

Early researchers of human social networks from August Comte to Jacob Moreno and contemporary researchers such as Mark Granovetter or Robin Dunbar enable us to understand the rich relationships that influence friendship patterns, scientific team collaborations, or international diplomatic conflicts. As the complex relationships become more visible and understandable, measurement and prediction become more reliable.

Jen Golbeck's lucid and insight-filled book makes a substantial contribution to explaining these modern phenomena as they play out on the social web. She

deftly integrates mathematical concepts with visual presentations, all conveyed with potent examples that engage and motivate readers.

For the first time in history, much of what we do is mediated electronically, and for the first time in history we are developing the tools, shown in this book, to make social behavior patterns visible. This growing capacity to perceive, understand, measure, and predict brings enormous power to those who master these network analysis skills. Measuring relationships and seeing changes over time is the first step to predicting future performance. More importantly network analysts gain the power to make bold decisions and take effective actions that influence outcomes in communities, markets, health/wellness, environmental preservation, sustainable energy, and many more domains of human activity.

Technology-mediated social participation is a rapidly rising force in which the chain reactions of human collaboration can overthrow oppressive regimes, influence democratic elections, and shape economic successes. These chain reactions can also trigger cascades of human activity that reduce obesity, support smoking cessation, encourage energy conservation, and accelerate citizen science.

However, social media can also be used by oppressive regimes to track/suppress opponents, extremists to promote racial hatred, or terrorists to coordinate their attacks. Since malicious spammers, criminal gangs, and illegal traffickers can also use these potent technologies, researchers and policy makers are well advised to develop strong skills in responding with pro-social strategies that protect the public.

Social media developers could soon find themselves with ethical dilemmas similar to those faced by nuclear physicists in the late 1940s. Having developed a potent technology, the dangers of misuse could threaten the huge positive opportunities, which could bring stunning benefits for future generations.

Realizing the benefits of the social web is a grand interdisciplinary project that will play out over many decades and require new skills and substantial contributions from a wide variety of disciplines that span computing sciences, social sciences, communications, and more. The benefits will accrue most to those individuals, organizations, disciplines, and nations that appreciate the potential and take action.

This visionary book and many more are necessary to educate a new generation of students, researchers, and policy makers, so that they can perceive, understand, measure, and predict future directions. More importantly, they will be able to intervene to produce more positive outcomes.

Ben Shneiderman

Preface

The web has always been fast-growing, but for a decade, it was mostly a place where users only read content. Now, social features are present on many websites, and understanding users' interactions is a complex and far-reaching topic. To cover all the interesting questions and methods of analysis would require volumes of text. At the core of all of these complex interactions are relationships that people have online, both directly with other people and through the content they create. The goal of this book is to introduce techniques for analyzing those social relationships.

Online relationships form rich networks, and many fields of study have methods for analyzing them. Math, computer science, sociology, biology, information studies, business, and others have solid, systematic methods for understanding networks in one form or another. To really understand the networks found on the social web, tools from all these traditions are necessary.

Until now, no textbook had been published that integrated all these approaches. An instructor trying to teach a course on analyzing social media networks was forced to either focus on only one or two related types of analysis or to assemble readings from many diverse sources. Although the latter approach has the benefit of a broader scope, it can lead to redundancy and a loss of context around ideas. That makes it harder to understand the background, motivation, and application of each technique.

I experienced this problem in my own classes that I have taught on this subject. After several semesters of trying different approaches, I found the best solution was to write my own text for the class, which I supplemented with other readings for each lesson. This worked better for students and made it easier to teach.

The job market increasingly demands that students have expertise in analyzing social media, and social networks and social media are hot areas of research across academia. As a result, more and more courses are being created in universities to teach these skills. Based on my experiences teaching these courses, I decided to write this book.

This book is organized by idea, not by discipline. It starts with basics about network structure, and it includes ideas from many backgrounds on that topic. Each chapter adds a new type of analysis, which may include techniques from many areas of science, social science, and humanities research. The goal is to give students a broad set of tools with which they can understand the networks before them.

Much of the book follows a research-based teaching model. The first half is dedicated to teaching analysis techniques, and the second half uses case studies to show how the analysis can be applied in different domains. Many exercises ask students to come up with their own questions about specific networks and to try

to answer them. I have found this helps students internalize the techniques and better understand when to use each (either alone or in combination).

And finally, since network analysis is so interdisciplinary, this book is not targeted at students from any particular background. I include some in-depth discussion of techniques from computer science, sociology, and epidemiology, but in a way that I hope is accessible to all readers. Students may choose to pursue these concepts more deeply within their individual areas of expertise, and there are many open-ended exercises designed to encourage this.

For students and instructors, we have created a companion website. It contains datasets, additional exercises and project ideas, and tutorials on how to use the different tools and analysis techniques presented in the book. For instructors, we also provide lecture outlines, slides, and solutions to selected exercises.

CHAPTER

Introduction 1

Social media has become the dominant method of using the Internet, and it has infiltrated and changed the way millions of people interact and communicate. Social networking in particular has become extremely popular, with over one billion users on Facebook alone and billions more accounts across thousands of social networking sites online.

Understanding social networks—both those explicitly formed on social networking websites and those implicitly formed in many other types of social media—has taken on new importance in light of this astounding popularity. Analysis of these social connections and interactions can help us understand who the important people are in a network, what roles a person plays, what subgroups of users are highly interconnected, how things like diseases or rumors will spread through a network, and how users participate.

Applications of these analyses are extensive. Organizations can prevent or control the spread of disease outbreaks. Websites can support participation and contributions from many types of users. Businesses can provide immediate assistance to customers who have problems or complaints. Users can band together to better understand their communities and government or take collective action. Content providers online can filter and sort information to show users the most relevant, interesting, and trusted content.

The methods for analyzing social networks have been around for decades or longer, but social media provides new challenges and opportunities. Networks online are orders of magnitude larger than the networks analyzed in the past. Often, the networks are simply too big to be analyzed in their entirety. A good social network analyst working with social media needs to know how to analyze the structure of networks, apply sociological principles to understand user behavior, and deal with the size, scope, and application of the networks.

This book is designed to teach the reader a range of social analysis techniques, how to apply them specifically to social media networks, and to illustrate a number of specific social media cases to which the techniques can be applied.

This chapter will present a history of the social web and an overview of the major types of analysis. For background, the types and details of websites used throughout the book will be covered, as well as some free tools that are useful for visualizing and understanding networks.

Analyzing the social web

Classic social network analysis studies a network's structure. In a social network, a person is considered a *node* or *vertex*, and a relationship between people is a *link* or *edge*. When all the people and relationships are identified, there are many statistics that can provide insight into the network. However, even before learning those statistics or anything about social network analysis, you can probably identify some important and interesting things in a network.

Consider Figure 1.1. Each circle is a person or node, and each line connecting them is a relationship or edge.

What things can you say about this network, without any training in social network analysis? We can see that node a has a lot of relationships. There is a long series of relationships from a to b to b1 to b2 and so on. There are many relationships among the nodes a1 through a10 in the lower right. That might be a group of people with very close relationships.

The first part of this book will introduce formal methods for quantifying these types of insights. This will include measures of a person's importance, how well connected the people in the network are, and which people form communities or clusters together.

These statistics are frequently used and often provide good insight into the nature of a network. However, those quantitative measures are not the only interesting ways to understand a social network. We will also look at qualitative attributes of the network. Tie strength, which is the strength of the relationship

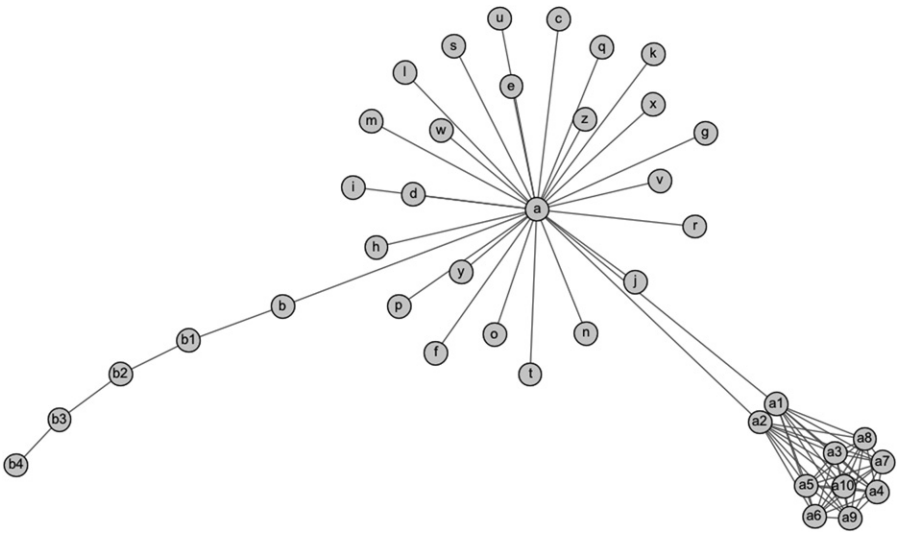


FIGURE 1.1 A sample social network

between two people, and trust are two relationship features that have great impact on what happens in a social network. Furthermore, learning what role a person plays in a network by analyzing his or her behavior can link quantitative measures with qualitative analysis to help better understand what goes on in a social group. Visualization, which is the creation images like Figure 1.1 that visually represent the structure of a social network, allows us to leverage our natural abilities to perceive patterns in images to better understand network structure and patterns.

With those analysis methods at hand, the next step is to use them to understand network phenomena. One of the most important of these phenomena is propagation: How do things like information, diseases, or rumors spread in a network? A combination of quantitative and qualitative features inform our understanding of propagation, and another set of analysis techniques is available to study the spread of things through networks.

Throughout the book, we will use real social media networks to demonstrate the techniques described above. But understanding social media goes beyond these types of analysis. The second half of the book will look at specific questions of interest to different types of social media. For example, what motivates people to contribute to Wikipedia? How do politicians leverage social media to spread their messages or communicate with constituents? How do businesses make decisions about when to use social media? What privacy threats do users face in these websites? To answer these questions, we will apply the techniques from the first half of the book and described above, and present the results of research and experiments to show the full range of analysis used to understand the many issues related to social media.

A brief history of the social web

The web was invented in 1991, and from the start, Tim Berners-Lee, its inventor, saw it as a place where people could interact. He called it "a collaborative medium, a place where we all meet and read and write." At first, authoring web content required people to learn HTML, the language used for all web pages. Putting pages online also required access to a server and some technical knowledge that was a barrier for casual web users.

There were some ways to interact—chat rooms and discussion forums existed even before the web—but overall, the web was a place of static web pages that users simply visited. Blogging began in 1997, and the website Blogger (now owned and operated by Google) went online in 1999. Not only did this allow users to generate content without any knowledge of HTML or other programming languages, but people could comment, thus allowing interaction online. Users could also follow each other's blogs, which created a social network behind the content.

The first site to launch in the spirit of modern social networking sites was Six Degrees. It went online in 1997 and allowed people to create profiles and list their friends. At the height of its popularity it had one million members.

Blogging and other interactive web technologies continued to grow through the millennium as the dot com era boomed and after the bubble burst. While some sites failed, some current major social media sites emerged. Friendster launched in 2002, which grew quickly and was the first major social networking website. It was followed by LinkedIn (a business-oriented network) and MySpace in 2003. MySpace was the social network that largely brought online social networking into the public consciousness, and it reigned as one of the most popular networks for several years. Facebook followed in 2004. It was first restricted to students at Harvard and a set of elite universities, but eventually expanded to all colleges and then the general public. It is currently the largest social networking website, with over a billion users.

Other social media technologies were coming online as well. In 2004, Flickr, a photo-sharing website, and Digg, a social bookmarking website, launched. YouTube, the video-sharing website, came online in 2005, and Twitter launched in 2006, introducing microblogging to the social media space.

At that point, most of the major technologies of social networking were up and running, but new developments still continued at a dramatic rate. Sites came online and failed every day, and successful sites' numbers of users grew at a dramatic rate. After the first few years of the millennium, social media was posing a challenge to the dominance of "traditional" web content. User-generated content from blogs, shared links, comments, forum posts, and social media content became more common than any other type of content, prompting *Time Life* Magazine to declare "You" as the person of the year in 2007.

While Google reigned as the most popular and most-used website for many years, Facebook surpassed it in 2010. Although varying from month to month, social media sites often make up at least half of the top ten most popular websites as tracked by Alexa.[1]

Websites discussed

The techniques in this book are not designed for any specific website or type of network; they are general techniques that will work on any network regardless of its source. We will consider networks built from all types of interactions and websites, from email to discussion boards, Facebook-style social networks to blogging, and including offline social networks drawn from people's behavior and even from literature. However, because the book is focused on social media, a number of popular sites and types of social media occur throughout the text. This section introduces those and provides some background.

Some of the most popular sites in 2012 will feature prominently in the book's discussions. Facebook is by far the largest of these. Launched in 2004, it has since

[1] http://www.alexa.com/topsites/.

grown to be the world's largest social network with over a billion users. It is a traditional social networking site, where users make explicit connections to "friends" and share updates with them. Other popular social networking sites include LinkedIn, which is geared toward professional relationships, MySpace, the social network that was most popular before the rise of Facebook, and Renren, a large social network based in China.

Twitter is another dominant website in the social media space, with 200 million active users in 2012. Twitter is called a microblog. Users post messages that are limited to 140 characters. It has social networking characteristics as well. Users can follow others they find interesting, and the posts, called "tweets," from anyone followed will appear on the user's main page. Unlike the case with many social networks, the relationship does not have to be mutual. If Alice follows Bob on Twitter, Bob does not have to approve the relationship or follow Alice back. Twitter is the main microblogging website in the United States, but Weibo in China is also extremely popular.

Twitter segues into a type of social media based on sharing certain types of information. Twitter lets people share short pieces of text, but many sites support sharing other types of media. Photo-sharing sites are popular, and one that will appear frequently in this book is Flickr. It allows users to post photos, label them with descriptive keywords called tags, and share them in a variety of ways. It also has a social networking component. Users can be friends with others, and this feature can be used to adjust access to photos. In addition, people can comment on the photos that others share, and this commenting behavior can also be used to form a social network. YouTube, which is owned and run by Google, is the most popular video-sharing website. Like Flickr does with photos, YouTube lets users upload, share, and comment on videos. They can also become friends with other users.

Social bookmarking sites allow users to share interesting links. Digg, del.icio.us, and Reddit are popular sites for this activity. They support tagging links, voting them up or down to indicate interest. Pinterest is another social bookmarking site growing in popularity. It is visual, where users share photos that often link back to an originating article.

Tools used

Most of the techniques you will learn in this book require no special software and no complex calculations. However, to compute statistics about every node in a network can be time consuming, and some methods are too complex to apply by hand. A number of tools are available that will help with social network analysis, and two in particular are discussed in this book. They are free, have many built-in methods for assisting with social network analysis, and have easy-to-use user interfaces for creating visualizations of networks and interacting with them.

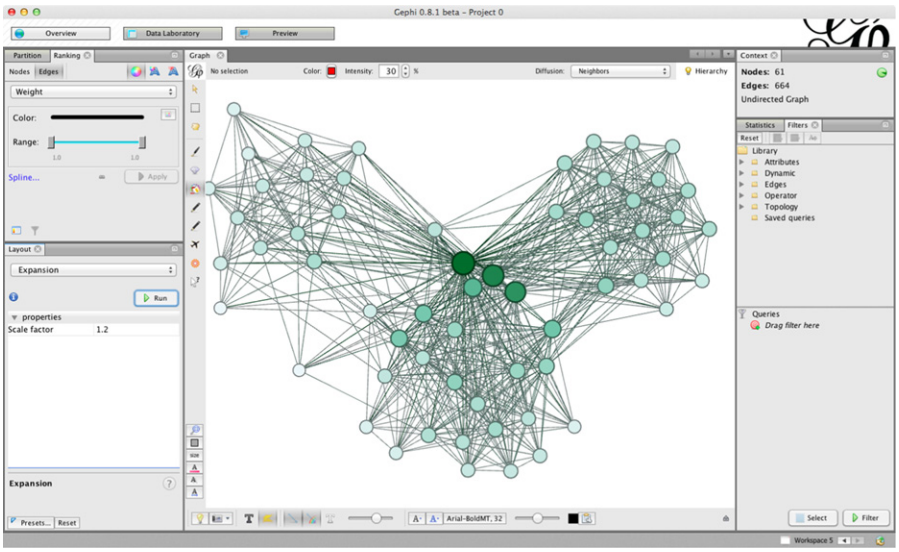


FIGURE 1.2 The Gephi interface

The first is Gephi (Figure 1.2).[2] It is an open-source free software package that runs in Windows, Mac OS X, and Linux. Gephi is a visualization tool with capabilities to calculate centrality, clustering, network diameter, and other metrics. Because it is open-source, there are also many plugins that add functionality to the core program.

The second tool is NodeXL (Figure 1.3), a template for Microsoft Excel 2007, 2010, and later Excel versions on Windows. It is a free download. Like Gephi, it has tools for visualizing graphs and computing many common network analysis statistics.

Both tools have features called spigots which allow users to directly import network data from other sources. Gephi comes with an email network importer, but other spigots are available as plugins. NodeXL can import email as well as queries to Twitter, Flickr, and YouTube. These spigots make it easy to get network data for analysis and experimentation.

Exercises

1. (Group exercise) List all the social media websites you can think of that may be turned into a social network. Try to group these sites thematically. What features do they share? How are they different?

[2] http://gephi.org.

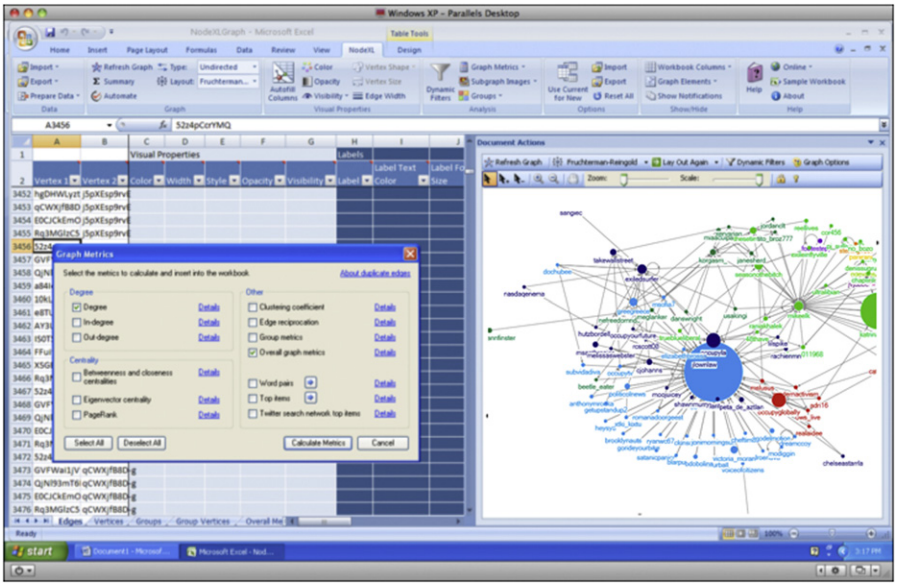


FIGURE 1.3 The NodeXL interface

2. What are some ways you think social network analysis can be useful? Before you have learned the analysis techniques that will be covered in this book, explain what you think you might be able to accomplish with a better understanding of who is connected with whom.
3. The terms social network and social media are used frequently in this text. What do you think each means? What are the relationships, similarities, and differences?
4. Think about the social media you use from your list in #1.
   a. What sites do you personally use?
   b. Do they have overlapping features (e.g., do you use a photo-sharing website in addition to Facebook, which allows you to share photos)?
   c. If so, why do you use two sites instead of one?
   d. If you do not use sites with overlapping features, is it a conscious choice to keep your social media content consolidated, is there another reason, or have you not thought about it? Would you ever consider using sites with overlapping features? Why or why not?

This page intentionally left blank

CHAPTER

Nodes, Edges, and Network Measures

2

The term *social network* has entered common language and is understood to describe circles of friends, acquaintances, colleagues, and so on. However, networks are well grounded in mathematics, and understanding how to represent, describe, and measure properties of networks will be the foundations of quantitative network analysis.

Basics of network structure

"Six degrees of separation"—the idea that people who seem very unlike one another may be connected by a chain of six or fewer mutual acquaintances—is one of the major network analysis concepts that has made its way from academic research (which will be discussed later in this book) into popular culture. It also is responsible for spawning the idea of "six degrees of Kevin Bacon" and the Kevin Bacon Game. The goal of the game is to connect any actor to Kevin Bacon through co-stars in movies, in as few steps as possible. For example, Elvis Presley was in *Live a Little, Love a Little* with John Wheeler, who was in *Apollo 13* with Kevin Bacon. This gives Elvis a "Bacon Number" of 2, since the path length (i.e., number of hops) between them is 2. The Bacon Number is a variation on the more well-established *Erdos Number*, a similar notion among mathematicians and computer scientists that maps how many co-author relationships separate them from the famous and prolific mathematician, Paul Erdos, one of the founders of graph theory used to analyze networks.

We will use the Kevin Bacon game throughout this chapter to understand some of the basic ideas of how networks are put together.

The network of actors in the Kevin Bacon game is one example of a social network. Each actor is connected to his or her co-stars. Consider the movie *Apollo 13* and its five stars: Tom Hanks, Gary Sinise, Ed Harris, Bill Paxton, and Kevin Bacon. To create a network of those actors, they can be linked if they were in another movie together.

We can represent each actor as a box. In the network these will be called *nodes* or *vertices*. The two terms can be used interchangeably, and both will appear depending on the background of the network analysis work you are reading. We will use *nodes* as the main term throughout this book.

The next step is to link the actors who have been in other films together. Tom Hanks and Gary Sinise were both in *The Green Mile*, so we can add a connection

between them. Tom Hanks, Bill Paxton, and Gary Sinise were all in a documentary called *Magnificent Desolation: Walking on the Moon* together. Thus, we can link each of them to one another. Gary Sinise was in *The Human Stain* with Ed Harris. He was also in *Beyond All Boundaries* with Kevin Bacon and Tom Hanks. In the network, we represent these links as lines that connect the actors. These may be called *links*, *edges*, and sometimes *ties.* We will primarily call them *edges* in this book.

A *network* or *graph* is a set of nodes and edges.

Knowing the nodes and edges is all that is needed to analyze a social network. However, edges can have a number of additional features, which can be used in analysis.

Edges can be *labeled*. The label describes something about the relationship between the people. It could name the relationship (e.g., sister, mother, cousin), or some information about the relationship. In Figure 2.3, the labels indicate a movie that each pair of actors have been in together.

Edges can be *weighted* or *valued*. We will use *weighted* in this book. The weight is a number that indicates numerical information about a relationship. Often, this is the strength of a relationship, but it can come from a variety of sources and indicate many things. In the *Apollo 13* example, we could weight the edges by the number of movies the actors have been in together. For all the edges in this graph, the weight is 1, because they have only been in one movie with each other (in addition to *Apollo 13*), except the edge between Gary Sinise and Tom Hanks. Since both actors were in *Magnificent Desolation* with Bill Paxton and in *Beyond All Boundaries* with Kevin Bacon, we know they were in at least those two films together. They were also co-stars in *Forrest Gump* and *The Green Mile*, so the weight between them would be 4. Weights can be shown as numeric labels on the edges, or the edges can be drawn thicker to show the greater weight. Figure 2.4 shows both of these options.

Edges can also be either *directed* or *undirected*. An undirected edge indicates a mutual relationship, whereas a *directed* edge indicates a relationship that one node has with the other that is not necessarily reciprocated. The type of edge used defines the network as either a *directed network* or an *undirected network*.

The *Apollo 13* example we have been following is an undirected network. It is undirected because if two actors are in a movie together, there is no notion of a one-way relationship. If Tom Hanks is in a movie with Gary Sinise, then Gary Sinise must also be in that movie with Tom Hanks. Undirected edges are drawn as simple lines between nodes, as is seen in Figures 2.1–2.4.

If we were to build a network of email communication, then we could have a directed network. Person A may send an email to Person B without receiving a reply. In that case, we would draw an edge indicating the one-way relationship from A to B, but not the reverse. In a directed network, edges can be reciprocated. Person A may email Person C, and C may reply. In this case, we want the edge to indicate a relationship in both directions. When showing a directed network, edges have arrowheads to show the direction of the relationship. If there is

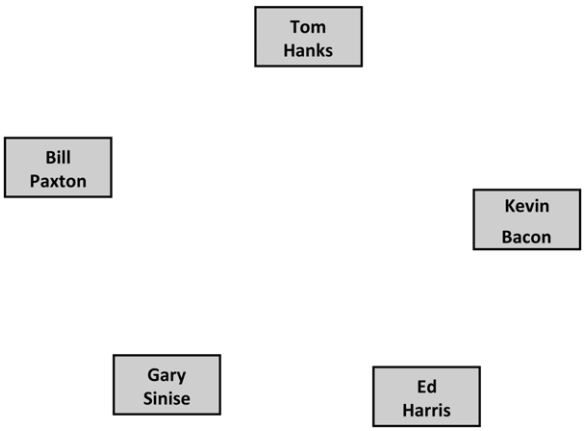


FIGURE 2.1

The five co-stars of *Apollo 13*. Each is represented as a node in the network.

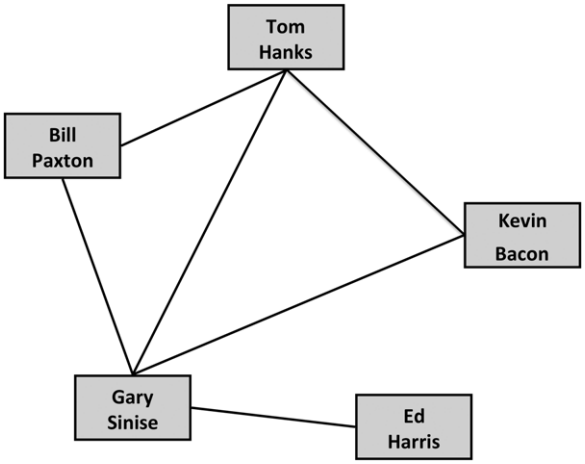


FIGURE 2.2

The edges connect actors who were in movies together.

a relationship that is reciprocated, it is either drawn with a line that has arrows on both ends, or by two directed edges as shown in Figure 2.5. An undirected edge is never used in a directed network.

In a directed graph, the number of possible edges is double that of an (otherwise identical) undirected graph. Between any two nodes, there can be only one edge in an undirected graph, but two possible edges in a directed one.

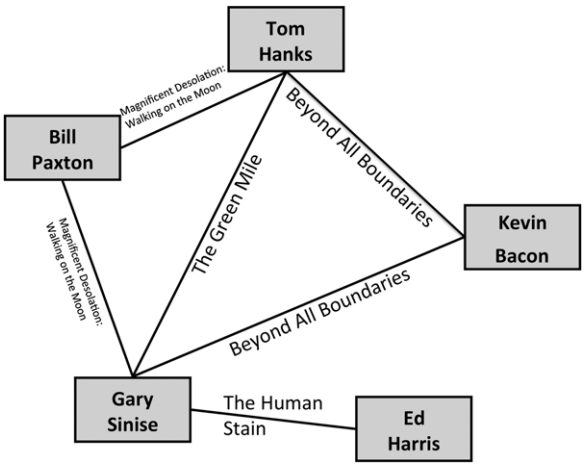


FIGURE 2.3

A labeled graph where the edges indicate at least one movie that the actors have been in together, not including *Apollo 13*.

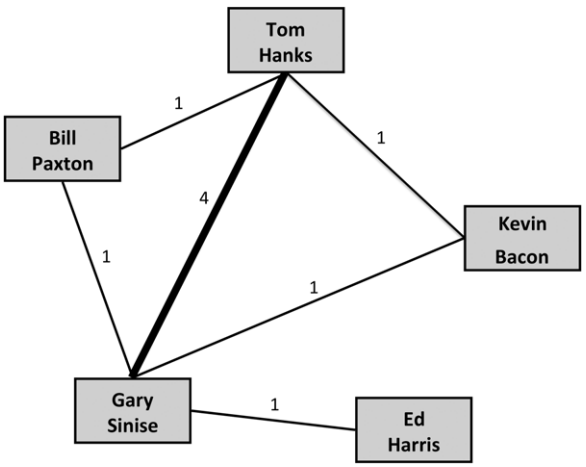


FIGURE 2.4

A weighted graph where weights are indicated both as numbers and by the thickness of the edge. In this graph, weight indicates how many movies the actors have been in together.

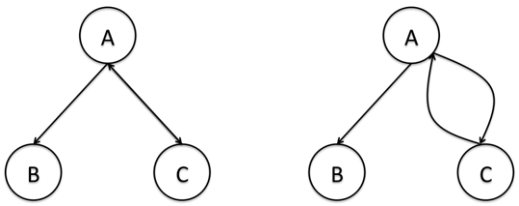


FIGURE 2.5

Two ways of drawing a *directed* network. The edge from A to B is directed only one way. The edge from A to C goes in both directions and can be drawn either as one edge with two arrow heads (left) or as two edges pointing in opposite directions (right).

Representing networks

The example networks we have seen so far are presented as figures with nodes represented as circles or squares and edges as lines that connect them. There are a variety of methods for representing networks. We will discuss more sophisticated visual techniques in Chapter 4: Network Visualization. In this section, we will focus on text-based representations. These are used as the inputs to many visualization techniques and are also necessary for graphs of most size since they quickly become too large to easily draw as we have in the figures above.

Adjacency lists

An *adjacency list*, also called an *edge list*, is one of the most basic and frequently used representations of a network. Each edge in the network is indicated by listing the pair of nodes that are connected. For example, the adjacency list for the *Apollo 13* network is as follows:

```
Tom Hanks, Bill Paxton
Tom Hanks, Gary Sinise
Tom Hanks, Kevin Bacon
Bill Paxton, Gary Sinise
Gary Sinise, Kevin Bacon
Gary Sinise, Ed Harris
```

Each line contains one pair of nodes. In this example, the names of the nodes are separated by commas, but you could also use tabs or other characters as a separator.

The order of these lines does not matter since there is no concept of order in networks. For example, we could move all the pairs with Tom Hanks to the end of the list, and the list would still accurately list all pairs of nodes that are connected by edges.

Since this network is undirected, the order of the node names in each pair is irrelevant, too. The current list has "Tom Hanks, Bill Paxton" as the first entry, but it would have the same meaning if we reversed the order to "Bill Paxton, Tom Hanks." However, if the network is directed, this would not be true. In a directed network, the order of the node names is important. If a pair is listed as "Node A, Node B" in a directed network, it means there is a relationship from Node A to Node B. The reverse relationship is not implied, but it can be indicated by including another line listing "Node B, Node A." If both pairs are listed, it means there is a relationship in both directions.

Adjacency lists can also include additional information about the edges, as was discussed in the previous section. This is included on the same line as the two node names, and usually follows them. An edge weight is a common value to see included in an adjacency list. Again using the Apollo 13 example and the edge weights from Figure 2.4, the list would be written as follows:

Tom Hanks, Bill Paxton, 1
Tom Hanks, Gary Sinise, 4
Tom Hanks, Kevin Bacon, 1
Bill Paxton, Gary Sinise, 1
Gary Sinise, Kevin Bacon, 1
Gary Sinise, Ed Harris, 1

Edge labels can also be included in an adjacency list in the same way.

Adjacency matrix

An alternative to the adjacency list is an *adjacency matrix*. In an adjacency matrix, a grid is set up that lists all the nodes on both the X-axis (horizontal) and the Y-axis (vertical). Then, values are filled in to the matrix to indicate if there is or is not an edge between every pair of nodes. Typically, a 0 indicates no edge and a 1 indicates an edge.

The Adjacency Matrix for the Apollo 13 Network

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	1	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	1	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

Notice a couple of things about this matrix. First, the diagonal is all zeroes because there are no edges between a node and itself in our example. Some networks do allow for self-loops. For example, in an email network, if a person emails himself, there could be a link from one node to itself, and thus there would be a 1 on the diagonal. Second, the matrix is symmetric. The numbers in the first row are the same as the numbers in the first column. The numbers in the second row are the same as the numbers in the second column. This is because the graph is undirected. Just as in the adjacency list, where the order of pairs in an undirected graph didn't matter,

Notice that the Diagonal, Indicating a Person's Link to Himself, is all 0s

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	1	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	1	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

If we have a directed network, the matrix will not necessarily be symmetric. For example, consider the small network in Figure 2.5. In this case, there are edges from A to C, and C to A, and from A to B, but the reciprocal edge from B to A is absent. Thus, we only record a 1 for the A−B edge, and record a 0 for the B−A edge. The adjacency matrix would look like this:

A Small Adjacency Matrix for a Directed Network

	A	B	C
A	0	1	1
B	0	0	0
C	1	0	0

In the examples we have seen so far, we have been recording a 1 in the matrix to indicate an edge is present, and a 0 when there is no edge. This scheme can be altered to show the weight of an edge as well. To do this, we replace the 1 with

the edge weight. Using the values from Figure 2.4, we would have a weight of 4 between Tom Hanks and Gary Sinise. The matrix would look like this:

The Adjacency Matrix for the Apollo 13 Network with Edge Weights

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	**4**	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	**4**	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

XML and standard formats

In addition to the formats above, a common way to share network data is through standard formats like XML. XML, the eXtensible Markup Language, is the basis for many things on the web, including HTML—the language used to write web pages. It is a simple text format designed to be readable by any programming language on any operating system.

An example of how one might represent part of our example network in XML is as follows:

```
<Person>
   <name>Tom Hanks</name>
   <connection>Bill Paxton</connection>
   <connection>Gary Sinise</connection>
   <connection>Kevin Bacon</connection>
</Person>
```

The text contained between the <and> signs are tags. There are opening or "start" tags (e.g., <Person>) and then corresponding end tags that include a leading forward slash. These indicate the end of the section (e.g. </Person>). In this snippet of code, we are describing a "Person." The opening tag indicates that our description has started. Between the start and end tags, we list attributes of our person. To do that, we have more tags that describe attributes of the person. This example includes a name (between start and end "name" tags), and the person's connections.

XML can be far more complex than this, but this simple example shows the general structure. Instead of listing pairs of names like we would in adjacency list, connections are represented using XML tags. The benefit of XML is that it is easy to process, and many social network analysis tools are able to read in XML-formatted documents to load a social network.

There are a number of standard ways to describe social networks in XML. In the example above, we had a tag for "Person" and tags for "name" and "connection." The XML standards for describing social networks prescribe a set of tag names to use for describing social connections. Examples of these standards include GraphML (the Graph Markup Language) and FOAF[1] (Friend Of A Friend).

XML is not the only standard way to represent social networks. Other web formats, like JSON (JavaScript Object Notation), can be used to describe networks. These standards and formats are constantly evolving. As the amount of social network data online continues to grow and as organizations find more use for it, there will likely be updates to existing standards and new ones introduced.

Basic network structures and properties

Beyond nodes and edges, there are some basic structures that are important to know for describing and understanding networks. These include descriptions of nodes, their connections, and their role in the network.

Subnetworks

So far, we have considered the entire graph or network, looking at how many nodes and edges it has and how to describe them. Often, there are parts of the network that are interesting as well. When we are considering a subset of the nodes and edges in a graph, it is called a *subnetwork*.

Some of the simplest subnetworks are *singletons*. These are nodes that have no edges. While these nodes are not very "social," they are still part of a social network. In fact, it is very common to find singletons in online social networks. Often, these represent people who signed up for an account to access some part of the site other than the social networking features, or people who signed up but never actively participated. In Figure 2.6, node A is a singleton because it isn't connected to any other node in the network.

We also are interested in small groups of nodes. When looking at two nodes and their relationship, it is called a dyad, and a group of three nodes is called a triad. Figure 2.6 shows a connected dyad between B and C, and a fully connected triad between D, E, and F. However, we could consider the relationship between A and B. Even though they are not connected, that pair of nodes could also be called a dyad.

Cliques

Groups of nodes of any size have properties that are interesting. One of particular interest is whether or not all nodes in a group are connected to one another. When this happens, it is called a *clique*. The term is the same as the one we use

[1]FOAF is actually a Semantic Web standard and, while it is commonly presented as XML, it can be used in other standard formats. For more details, check out http://xmlns.com/foaf/spec/.

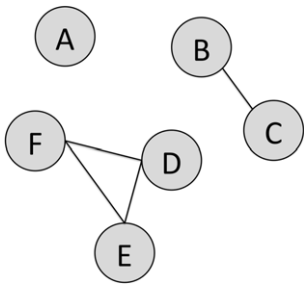


FIGURE 2.6 A social network with a singleton, dyad, and triad

to refer to, for example, a group of people who are all strongly connected and tend to talk mostly to one another (e.g., "Alice is part of a clique at school"). For a graph or subgraph to be a clique, every node must be connected to every other. In Figure 2.6, nodes D, E, and F form a clique. However, if the edge from D to E were missing, it would not be a clique.

Clusters

We are also interested in clusters of nodes. In Figure 2.6, we see a group of nodes to the lower right that have many connections between them. This group is not a clique because every node is not connected to every other. For example, node D is not connected to O and F. However, the group is clearly more connected to one another than the graph is as a whole or compared to other subgraphs. While there is no strict definition of a cluster like there is for a clique, we can describe properties of clusters using some network measures, like density, that we will discuss later in this chapter. There are a variety of methods to automatically identify clusters based on the network structure.

Egocentric networks

One of the most important types of subgraphs we will consider is the egocentric network. This is a network we pull out by selecting a node and all of its connections. In Figure 2.6, node D is connected to nodes A, E, B, C, and Q. There are edges from D to each of these nodes and edges between them. When considering egocentric networks, we can choose which of those to include. Consider Figure 2.7.

Figure 2.8(a) shows Node D and its edges to its neighbors. Because we are going one step away from D in the network, this is called a degree-1 egocentric network. It only shows us the nodes D is connected to. More frequently, we want to know about the connections between D's neighbors.

If we want to see only D's neighbors and their connections, it is called a 1.5-degree egocentric network, shown in Figure 2.8(b). It is 1.5 instead of 2

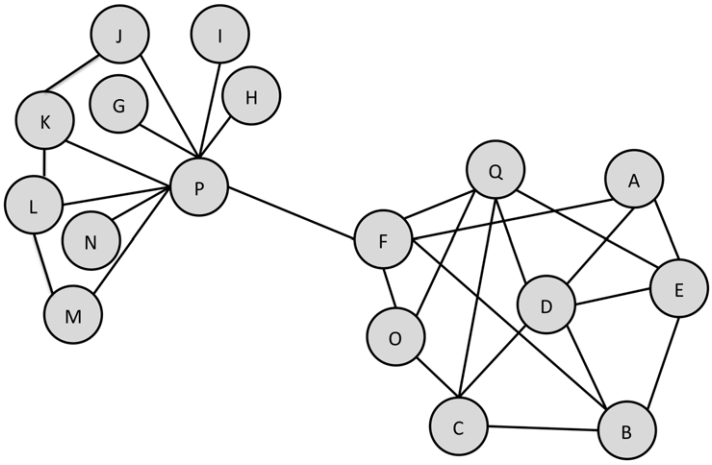


FIGURE 2.7 A sample undirected network

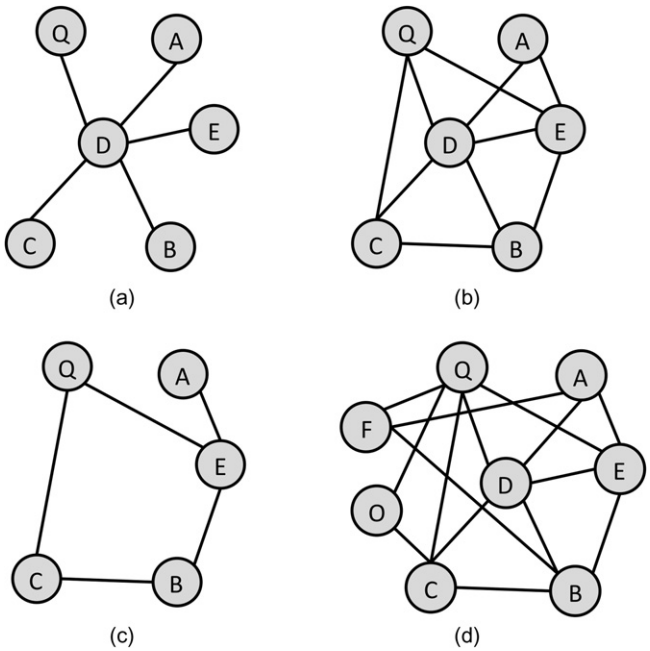


FIGURE 2.8

(a) The 1-degree egocentric network of D, (b) the 1.5-degree egocentric network of D, (c) the 1.5 egocentric network of D with D excluded, and (d) the 2-degree egocentric network of D.

because we are not going twq full steps away from D in the network. We are going only one step, but then looking at the connections between those nodes. However, including D in the graph is a bit redundant because we know that D is connected to all of the other nodes. Often, the central node and its edges are excluded and only the node's neighbors and there connections are considered, as in Figure 2.8(c). This helps make the graph more readable.

Egocentric networks can extend out further. Figure 2.8(d) shows the 2-degree egocentric network. It includes all of D's neighbors, their connections to one another, and all of their neighbors.

Egocentric networks are used to understand nodes and their role in the network. Egocentric networks are an important tool for network analysis throughout this book.

Paths and connectedness

The connections between nodes and measures of their closeness are important network characteristics we will discuss in this book.

Paths

A *path* is a series of nodes that can be traversed following edges between them. In Figure 2.7, there is a path connecting node M to node C by following the edges from M to P to F to O to C. To determine the length of a path, we count the number of edges in it. The path from M to C has a length of 4 (M−P, P−F, F−O, and O−C). There are longer paths from M to C. For example, we could follow M−L−K−J−P−F−Q−D−C. However, we are typically only interested in the *shortest path* from one node to another. Note that there may be multiple shortest paths between two nodes. In Figure 2.7, there are two shortest paths from Node F to Node E: F−A−E and F−B−E. Shortest paths will be an important measure we consider in network analysis and are sometimes called geodesic distances.

Connectedness

Paths are used to determine a graph property called *connectedness.* Two nodes in a graph are called connected if there is a path between them in the network. There does not need to be a direct edge, though that would count. Any path through a series of nodes will work. An entire graph is called connected if all pairs of nodes are connected.

In an undirected graph, this is relatively straightforward. A path is found by following edges between nodes. In a directed graph, edges may only go in one direction. Thus, while there may be a set of edges that connect two nodes, those edges may not all point in the right direction. If there are edges that can be followed in the correct direction to find a path between every pair of nodes, the directed graph is called *strongly connected.* If a path cannot be found between all pairs of nodes using the direction of the edges, but paths can be found if the

directed edges are treated as undirected, then the graph is called *weakly connected.*

If a graph is not connected, it may have subgraphs that are connected. These are called *connected components.* For example, Figure 2.6 includes a three-node connected component, a two-node connected component, and a singleton.

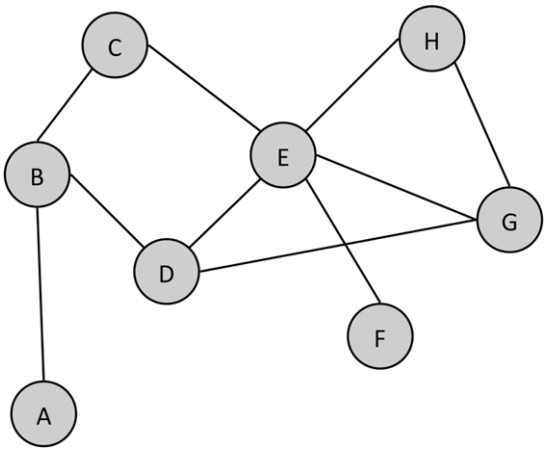
Bridges and hubs

We will discuss many ways of determining the importance of edges and nodes later in the book when discussing Centrality. However, there are two basic concepts that we can use to identify particularly important edges and nodes right off.

The first is a *bridge*. Intuitively, a bridge is an edge that connects two otherwise separate groups of nodes in the network. Formally, a bridge is an edge that, if removed, will increase the number of connected components in a graph. In Figure 2.7, the edge between nodes P and F is a bridge because if you take it out, the group of nodes on the right will be totally disconnected from the group of nodes on the left.

Hubs are important nodes rather than edges. They do not have a definition as strict as that of a bridge, but the term is used to refer to the most connected nodes in the network. In Figure 2.7, node P would be a hub because it has many connections to other nodes.

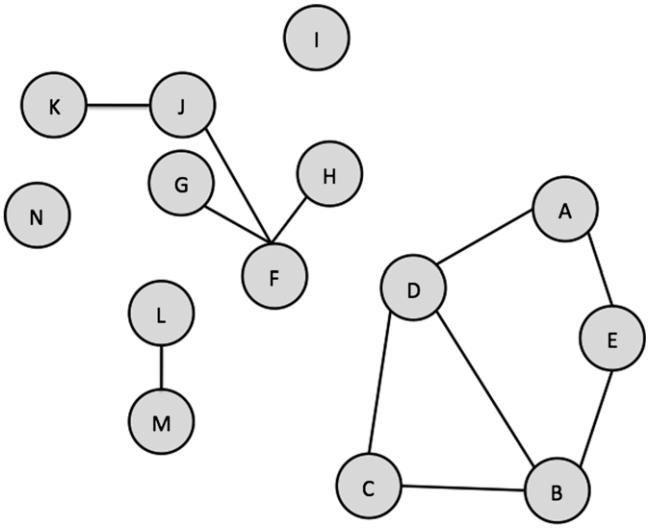
Exercises



1. Answer the following questions about this graph.
   - **a.** How many nodes are in the network?
   - **b.** How many edges are in the network?
   - **c.** Is this graph directed or undirected?

d. Create an adjacency list for this graph.
e. Create an adjacency matrix for this graph.
f. What is the length of the shortest path from node A to node F?
g. What is the largest clique in this network? How many cliques of that size are there?
h. How many connected components are there in this network?
i. Draw the 1.5 ego network for node E (without including node E in the graph). How many singletons are in the ego network?
j. Are there any hubs in the network? If so, which node(s) and why is it a hub?

2. Consider this graph



a. How many singletons are there in the network? List them.
b. What is the largest connected component?
c. Are there any bridges in the network? If so, where are they?
d. Create an adjacency list for the network.
e. Create an adjacency matrix for the network.

3. List at least three different networks that exist within Facebook. For each one, answer the following:
a. What constitutes a node?
b. What constitutes an edge?
c. Is it directed?
d. Is it weighted? If so, what does the weight indicate?
e. What is the smallest component in the graph?

4. List the 15 people you are closest to. Turn this list into a network by listing all the connections between these people.
a. Is your network directed or undirected?

   b. What do the edges represent (friendship, family relationship, close relationships, acquaintances, etc.)?
   c. Give the adjacency list for the network.
   d. Give the adjacency matrix for the network.
   e. Are there any singletons?
      i. What is the largest clique?
      ii. Are there any bridges? If so, where are they?
      iii. Are there any hubs? If so, which nodes are hubs and why?
      iv. How many connected components are there in the graph?

5. Repeat exercise 4, but instead of listing the 15 people you are closest to, choose 15 people with whom you only have a casual relationship—co-workers, classmates, and other acquaintances.
   a. Repeat all the subparts of exercise 4 for this network.
   b. Compare the results you obtained from the two graphs. Where are there big differences? Why do you think this is?

This page intentionally left blank

CHAPTER

Network Structure and Measures

3

In the previous chapter, we covered the basic vocabulary of networks. This chapter will cover methods for understanding and comparing networks, the role of nodes within networks, measuring importance, and related properties.

The properties introduced here and the measures for quantifying them are important for many network analysis tasks. They allow an analyst to identify important or influential individuals, characterize the network structure, understand how individuals fit within the landscape of the network, and carry these properties forward to understand how and why things happen in a network. The field of social network analysis has many well-developed and validated measures that allow us to systematically characterize networks and the nodes within them.

Describing nodes and edges

Many network properties describe how nodes are connected to one another and to the network as a whole. The simplest of these is *degree*. The degree of a node is the number of edges connected to that node. In undirected graphs, the degree of a node is simply the total number of edges connected to it. In directed graphs, there are two measures of degree: *in-degree* and *out-degree*. The in-degree is given by the number of edges coming into the node. In network diagrams, in-degrees are shown as edges with arrows pointing at the node. The out-degree is the number of edges originating from the node going outward to other nodes. These are shown with arrows pointing away from the node. The sum of the in-degree and out-degree gives you the total degree for the node.

Consider the graph in Figure 3.1. The degree of node P is 9. It is connected to nodes M, N, L, K, G, J, I, H, and F. The degree of node A is 3. It is connected to nodes F, D, and E.

Now look at Figure 3.2. This is a directed graph; some edges go in both directions, while others go in only one direction. In directed graphs, the in-degree and out-degree are both counted. The in-degree of node A is 2. It has edges coming in from nodes G and I. The out-degree is 3. It has edges going out to I, H, and L. Thus, the degree of node A is 5 (the sum of its in-degree and out-degree). Note that A's connection to I is counted twice, since the outgoing edge is considered separate from the incoming edge.

The out-degree of node N is 2, while the in-degree is 0. There are no edges coming in to N. The total degree for N is 2.

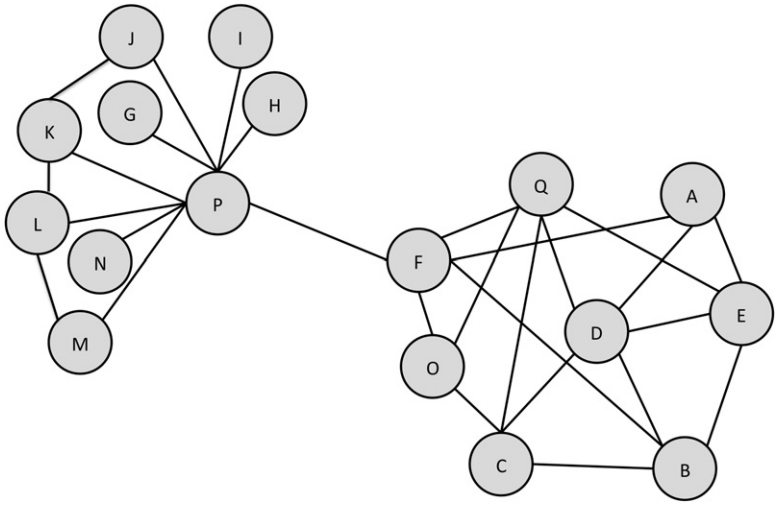


FIGURE 3.1

A sample undirected network.

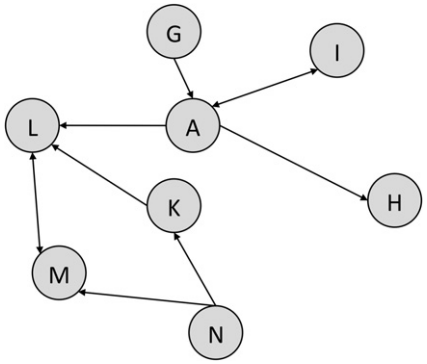


FIGURE 3.2

A sample directed graph.

Determining which nodes are most important or influential is the issue we will discuss in the next section on Centrality.

Centrality

Centrality is one of the core principles of network analysis. It measures how "central" a node is in the network. This is used as an estimate of its importance in the

network. However, depending on the application and point of view, what counts as "central" may vary depending on the context. Correspondingly, there are a number of ways to measure centrality of a node. In this chapter, four types of centrality are considered: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

In network analysis, one or more of these measures may be reported in order to gain a better perspective on the network. A node may appear highly central with one measure but have low centrality with another. That does not mean one measure is incorrect, though; they are simply different ways of describing nodes. The interpretation of the centrality measures is left to a human analyst.

For all of the centrality measures discussed below, it may be difficult to compare across networks. A very important node in a small network may have centrality measures that would seem unimportant in a larger network. This chapter introduces the basic ways of computing centrality, but they may need to be scaled to facilitate comparisons.

Also, the measures below are calculated for undirected, unweighted graphs. When working with directed or weighted networks, these measures require modification. This has significant implications for how the values are interpreted. Some of these issues will be discussed below with each measure.

Degree centrality

Degree centrality is one of the easiest to calculate. The degree centrality of a node is simply its degree—the number of edges it has. The higher the degree, the more central the node is. This can be an effective measure, since many nodes with high degrees also have high centrality by other measures. In Figure 3.1, node P has the highest degree centrality of 9. Meanwhile, node F has a relatively low degree centrality of 5. Many other nodes have that same centrality value or higher (e.g., node D has a degree centrality of 5).

Indeed, as an extreme counterexample, there may be a network with a very large, dense group of nodes that comprise the majority of the graph (this is sometimes called the *core* of the network), but far out from the core along a chain of low-degree nodes may lie one node that is connected to a large number of nodes with no other connections (this is sometimes said to be on the *periphery* of the network). This is illustrated in Figure 3.3. Such a node would have high degree centrality, even though it is distant from the core of the network and most of the nodes.

Degree centrality is a good measure of the total connections a node has, but will not necessarily indicate the importance of a node in connecting others or how central it is to the main group.

Closeness centrality

Closeness centrality indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network. Consider Figure 3.4.

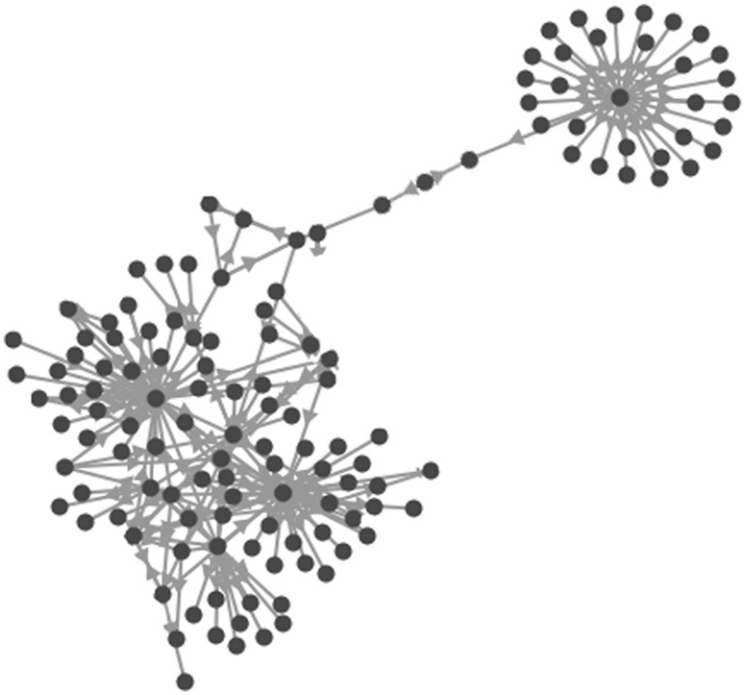


FIGURE 3.3

The node at the center of the cluster in the upper right would have a high degree centrality, even though it is far from the dense center of the network.

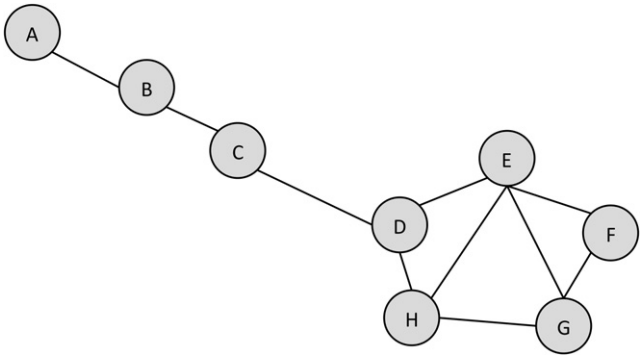


FIGURE 3.4

A sample network.

Let's start by computing the average shortest path length of node D. Table 3.1 shows each node and the length of the shortest path from D.

The average of those shortest path lengths is:

$$(3 + 2 + 1 + 1 + 2 + 2 + 1) \div 7 = 12 \div 7 = \mathbf{1.71}.$$

Note that we divide by 7 because there are seven other nodes.

Now repeat this for node A. This is shown in Table 3.2.

Here, the average shortest path length is:

$$(1 + 2 + 3 + 4 + 5 + 5 + 4) \div 7 = 24 \div 7 = \mathbf{3.43}.$$

In the case of closeness centrality, or average shortest path length, lower values indicate more central nodes. Thus, since node D's closeness centrality is 1.71 and node A's is 3.43, node D is more central by this measure.

The benefits of closeness centrality are that it indicates nodes as more central if they are closer to most of the nodes in the graph. This strongly corresponds to visual centrality—a node that would appear toward the center of a graph when we draw it usually has a high closeness centrality.

Table 3.1 The Shortest Path Lengths from D to each Other Node in the Network

Node	Shortest Path from D
A	3 (D–C–B–A)
B	2
C	1
E	1
F	2
G	2
H	1

Table 3.2 The Shortest Path Length from node A to Every Other Node in the Network

Node	Shortest Path from A
B	1
C	2
D	3
E	4
F	5
G	5
H	4

Betweenness centrality

Betweenness centrality measures how important a node is to the shortest paths through the network. To compute betweenness for a node N, we select a pair of nodes and find all the shortest paths between those nodes. Then we compute the fraction of those shortest paths that include node N. If there were five shortest paths between a pair of nodes, and three of them went through node N, then the fraction would be $3 \div 5 = 0.6$. We repeat this process for every pair of nodes in the network. We then add up the fractions we computed, and this is the betweenness centrality for node N.

For example, consider Figure 3.4. Let's compute betweenness centrality for node B. There are 10 pairs of nodes to consider: AC, AD, AE, AF, CD, CE, CF, DE, DF, and EF. Without counting, we know that 100% of the shortest paths from A to every other node in the network go through B, since A can't reach the rest of the network without B. Thus, the fractions for AC, AD, AE, and AF are all 1.

From C to D, there are two shortest paths: one through B and one through E. Thus, $1 \div 2 = 0.5$ go through B. The same is true for the shortest path from D to C. For the remaining pairs—CE, CF, DE, DF, and EF—*no* shortest paths go through B. Thus, the fraction for all of these is zero. Now we can calculate the betweenness for B:

$$4 \times 1 \text{ (A to all others)} + 0.5 \text{ (DC)} + 0.5 \text{ (CD)} + 5 \times 0 \text{ (all remaining pairs)} = 4 + 0.5 + 0.5 + 0 = \mathbf{5}$$

In contrast, the betweenness centrality of A is zero, since no shortest paths between D, C, D, E, and F go through A.

Betweenness centrality is one of the most frequently used centrality measures. It captures how important a node is in the flow of information from one part of the network to another.

In directed networks, betweenness can have several meanings. A user with high betweenness may be followed by many others who don't follow the same people as the user. This would indicate that the user is well-followed. Alternatively, the user may have fewer followers, but connect them to many accounts that are otherwise distant. This would indicate that the user is a reader of many people. Understanding the direction of the edges for a node is important to understand the meaning of centrality.

Eigenvector centrality

Eigenvector centrality measures a node's importance while giving consideration to the importance of its neighbors. For example, a node with 300 relatively unpopular friends on Facebook would have lower eigenvector centrality than someone with 300 very popular friends (like Barak Obama). It is sometimes used to measure a node's influence in the network. It is determined by performing a matrix calculation to determine what is called the *principal eigenvector* using the adjacency matrix. The mathematics here are more complicated than this book will cover, but the principles of eigenvector centrality are important and intuitive. Not

only is it used to determine influence in social networks, but a variant of eigenvector centrality is at the core of Google's PageRank algorithm, which they use to rank web pages.

The main principle is that links from important nodes (as measured by degree centrality) are worth more than links from unimportant nodes. All nodes start off equal, but as the computation progresses, nodes with more edges start gaining importance. Their importance propagates out to the nodes to which they are connected. After re-computing many times, the values stabilize, resulting in the final values for eigenvector centrality.

Most network analysis software packages will compute eigenvector centrality (and most other centrality measures as well), so it is not necessary to learn the intricacies of computing eigenvectors. However, understanding the general principles behind the measure is useful to decide when it is the right measure to use in analysis.

Describing networks

A number of measures can be used to describe the structure of a network as a whole. As discussed above, density is one of these. Density—the number of edges in the graph divided by the number of possible edges—is one of the most common ways of describing a network. However, other statistics provide different insights into network structure.

Degree distribution

Degree is used to describe individual nodes. To get an idea of the degree for all the nodes in the network, we can build the *degree distribution*. This shows how many nodes have each possible degree.

To create a degree distribution, calculate the degree for each node in the network. Table 3.3 shows the degrees for each node in the graph shown in Figure 3.1.

The next step is to count how many nodes have each degree. This is totaled for each degree, including those for which there are no nodes with that count. Table 3.4 shows the node count for each degree in this network.

The most common way to show a degree distribution is in a bar graph. The x-axis has the degrees in ascending order, and the Y-axis indicates how many nodes have a given-degree. For the data in Table 3.4, we would make a bar graph as shown in Figure 3.5.

Density

Calculating density

A node's connections say a lot about its role in the network. This goes well beyond the degree of a single node or the degrees of all nodes in the network.

Table 3.3 Degrees for each Node Shown in Figure 3.1

Node	Degree
A	3
B	4
C	4
D	5
E	4
F	4
G	1
H	1
I	1
J	2
K	3
L	3
M	2
N	1
O	2
P	9
Q	5

Table 3.4 The Degree Distribution for the Network in Figure 3.1. The First Column Shows the Degree, and the Second Column Shows How Many Nodes have that Degree

Degree	Number of Nodes
1	4
2	3
3	3
4	4
5	2
6	0
7	0
8	0
9	1

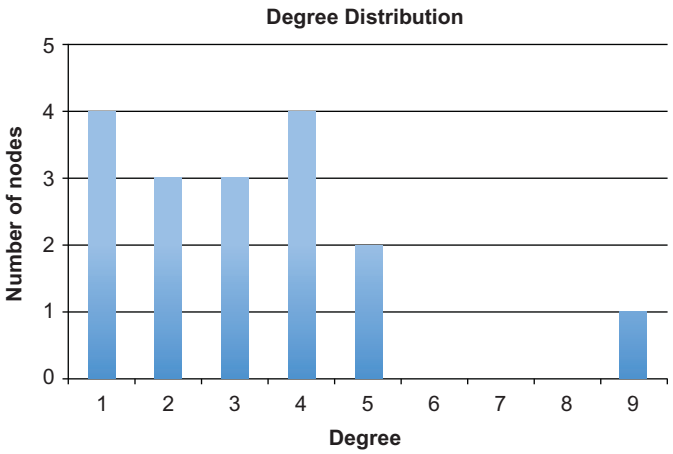


FIGURE 3.5

The degree distribution for the graph shown in Figure 3.1.

Another way to understand both individual nodes and the network as a whole is by studying *density*.

Density describes how connected a network is. More formally, it is a statistic comparing the number of edges that exist in a network to the number of edges that could possibly exist. Consider the following two networks, which both have the same number of nodes. Network (a) has very few edges while network (b) has numerous edges among the same number of nodes. Therefore, network (b) has higher density.

There is a formula to calculate density:

$$number\ of\ edges \div number\ of\ possible\ edges$$

The number of edges is something we can count in the network. The number of possible edges could also be counted by looking at each node and counting each of the other nodes that it could connect to. However, there is a simple formula for computing the number of possible edges as well.

First, consider the intuition behind the formula. If there are eight nodes in a network (as there are in the networks in Figure 3.6) each node can connect to seven other nodes. Node A can connect to B, C, D, E, F, G, and H. Node B can connect to A, C, D, E, F, G, and H. This scenario is sometimes known as the *handshake problem*—if a person comes into a room, how many people can he or she shake hands with? So if there are eight nodes in a network, and each node can connect to (shake hands with) seven others, then there are $8 \times 7 = 56$ possible edges. For a network with n nodes, we can generally say that there are $n \times (n - 1)$ edges. Each node can connect with every other node, excluding itself (hence the minus 1).

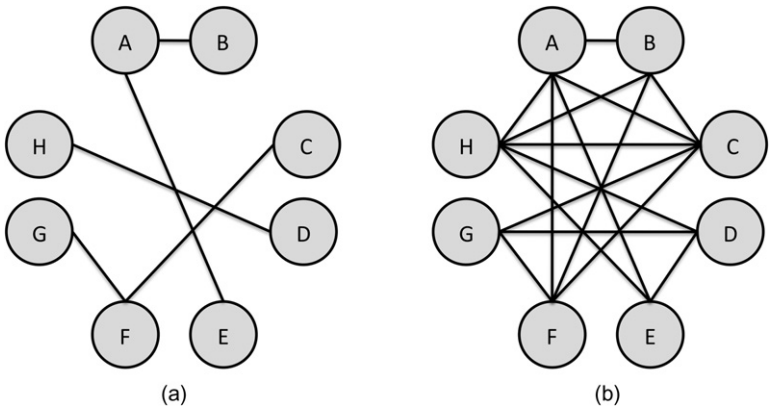


FIGURE 3.6

Network (a) on the left has fewer edges than network (b) on the right. Since they both have the same number of nodes and thus the same number of possible edges, network (b) is more dense.

However, it is not quite that simple. In this example, node A can connect to B and others, and node B can connect to A and others. Since each node can connect to 7 others, each connection has been counted twice. The connection from A to B is counted, as is the edge from B to A. In directed networks, this is fine—there are indeed two possible edges between A and B.

But in undirected networks (like the one in Figure 3.4), there can be only one edge between two nodes. Since the formula counts every node twice, simply divide by 2 to count the number of possible edges only once.

Thus, for **directed networks**, the number of possible edges in a graph with n nodes is:

$$n \times (n-1)$$

In **undirected networks**, the number of possible edges is:

$$\frac{n^*(n-1)}{2}$$

Now these formulas can be used to calculate density. In a directed network with n nodes and e edges, the formula for density is:

$$\frac{e}{n^*(n-1)}$$

In an undirected network with n nodes and e edges, the density formula is:

$$\frac{e}{n^*(n-1)/2}$$

We can use density to describe a network as a whole. Consider the networks in Figure 3.6. Both have eight nodes. Network (a) has five edges. Since it is an undirected network, the density is $(5/(8^*(8-1))/2)$ or $5 \div 28 = 0.179$. Network (b) has 16 edges, so the density is $16 \div 28 = 0.571$. Note that the density is higher for network (b), meaning it's denser.

A network with no edges would have a density of 0 (because the numerator in our equation would be 0, regardless of how many nodes there are). On the other hand, the densest possible network would be a network where all possible edges exist—a clique. As we just learned, the number of possible edges is the denominator of the density formula. In a clique, then, the numerator and denominator will be the same, so the density will be 1. This illustrates that density is always between 0 and 1, where 0 is the lowest possible density and 1 is the highest.

Density in egocentric networks

Density is a common way to compare networks. But it is even more commonly used to compare *subnetworks*—especially egocentric networks. Computing the density of each node's egocentric network gives us a way to compare nodes. Some will have dense egocentric networks, which means a lot of their friends know one another. Others will have sparse egocentric networks, and thus we know their connections often do not know one another. The density of an egocentric network is sometimes referred to as the *local clustering coefficient*.

To compute the density of an egocentric network, we use the 1.5-diameter network: We consider the node's connections and all the connections between those nodes.

For this calculation, the ego-node will be excluded from its egocentric network because the density of interest is that of the connections between the node's friends.

As an example, recall the network (b) from Figure 3.6. Node A is connected to nodes B, E, and H. To get the 1.5-diameter egocentric network, we will look at only nodes B, E, and H and the connections between them. This is shown in Figure 3.7(a). There are three nodes, so the number of possible edges is $3 \times 2 \div 2 = 3$.

Possible edges are from B to H and E (2) and from E to H (1)—a total of three. There are two edges in the network—from H to B and H to E. Thus, the density is 2 (the number of actual edges) ÷ 3 (the number of possible edges): $2 \div 3 = 0.667$.

The density of Node B's egocentric network can be computed from the 1.5-diameter egocentric network shown in Figure 3.7 (b). There are four nodes, so the number of possible edges is $4 \times 3 \div 2 = 6$. In the network, there are five edges (from A to H, F, and C, and additionally from C to F and H). So, the density of B's egocentric network is $5 \div 6 = 0.833$.

Thus, B's egocentric network (0.833) is more dense than A's (0.667). This is a common way to compare nodes in a network. However, having a higher egocentric network density does not necessarily mean a node is more "popular" or

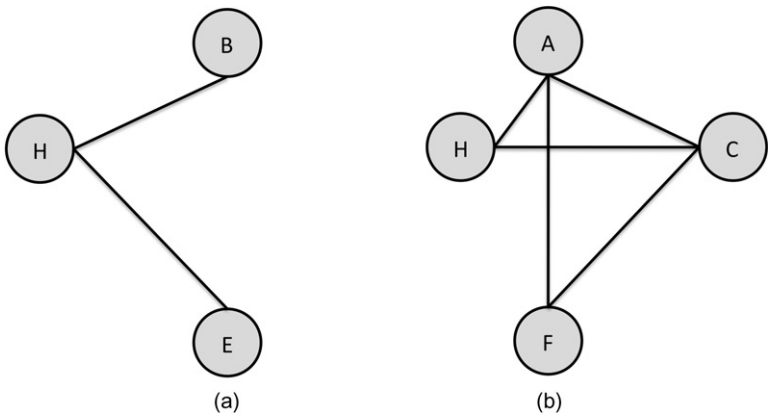


FIGURE 3.7

The 1.5-diameter egocentric networks for nodes A (a) and B (b) from Figure 3.2.

important. A node with a high degree (connections to many other nodes) will usually have a lower density. This follows the same logic we discussed above when comparing the density of small networks versus large networks. As the number of nodes in an egocentric network increases, the number of possible edges increases at that rate squared. Thus, more popular nodes tend to have lower densities.

Connectivity

Density measures the percentage of possible edges in a graph. *Connectivity*, also known as *cohesion*, measures how those edges are distributed. Connectivity is a count of the minimum number of nodes that would have to be removed before the graph becomes disconnected; that is, there is no longer a path from each node to every other node.

In Figure 3.4, the connectivity is 1 because removing node B, C, or D would disconnect the graph. Since removing any one of those nodes disconnects the graph, the connectivity is 1. In Figure 3.8, the connectivity is 2. Removing any one node would not break the graph into two parts, but there are several options for removing two nodes that would. For example, removing nodes E and F would separate G from the rest of the graph. If we removed B and D instead, node A would become separated.

Centralization

Centrality is an important way to understand the role of a node in the network and to compare nodes. *Centralization* uses the distribution of a centrality measure to understand the network as a whole. Any of the centrality measures presented

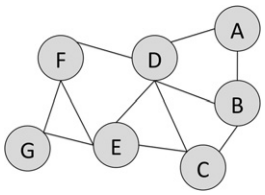


FIGURE 3.8

A sample network with a connectivity of 2.

above can be used, but only one is used at a time when computing centralization. If one node has extremely high centrality while most other nodes have low centrality, the centralization of the graph is high. If centrality is more evenly distributed, then the centralization of the network is low.

Centralization of power is an often-used concept and phrase, which relates very closely to centralization in a graph. For example, betweenness centrality can represent the control one node has in the ability of others to communicate. If many messages must pass through a particular node along their shortest paths, that node has the power to stop or pass on information. If a few nodes have very high betweenness, we can say that the power is centralized in those nodes.

Centralization is computed by looking at the sum of the differences in centrality between the most central node and every other node in the network, and dividing this by the maximum possible difference in centrality that could exist in the graph (Freeman, 1979). Since there are different centrality measures (e.g., betweenness, closeness, etc.), there are different centralization measures for a graph. But the basic formula is the same, and different centrality measures can be substituted.

Let $C(n)$ be the centrality of node n, using whatever centrality measure we choose. Say n^* is the most central node. We want to find the difference in centrality between n^* and every other node in the network, and add those up. If there are N nodes in the network, the formula for this is:

$$\sum_{i=1}^{N} C(n^*) - C(n_i)$$

Then, we want to divide this by the sum of the *maximum possible differences* between n^* and every other node. However, this maximum possible centrality will change depending on which centrality measure we are using. Denote this by using the same formula with *max* in front.

$$\max \sum_{i=1}^{N} C(n^*) - C(n_i)$$

Now, we can compute centralization. It is equal to the sum of the differences (the first formula) divided by the maximum possible sum of differences (the second formula):

$$\frac{\sum_{i=1}^{N} C(n^*) - C(n_i)}{\max \sum_{i=1}^{N} C(n^*) - C(n_i)}$$

Remember that to calculate betweenness, the fraction of edges from a node to each neighbor that go through the node in question are summed for every pair. In this case, for every pair of edge nodes, 100% (a fraction of 1 on a 0−1 scale) go through the center node. Thus, for each of the $n-1$ nodes, we have $n-2$ other nodes, so $1 \times (n-2)$ will be added to the centrality of the center node. This will be the case for all $n-1$ other nodes in the network. So, the maximum centrality difference is:

$$(n-1) \times (n-2) \times 1$$

Small worlds

If one phrase from social network analysis has made its way into common vocabulary, it is *six degrees of separation*. It is the title of both a play and a movie, and the origin for pop culture phenomena like the Kevin Bacon Game. The core idea behind the phrase, as we noted earlier, is that any two people in the world are separated by short paths, on average about six steps. Whether or not a network is a small world is an important property that relies on node and network measures described above.

Along with the notion of six degrees of separation"came the term *small worlds*, which indicates that people who may be very far apart physically and socially are still connected with relatively small paths. These ideas emerged long before their pop culture debuts. While such ideas were discussed since the early 1900s, the fundamental research on this topic was done in the 1960s by Stanley Milgram (Milgram, 1967).

Milgram wanted to explore the interconnection of social networks, so he devised an experiment. He sent information packets to people who lived in Omaha, Nebraska and Wichita, Kansas. The recipients were asked to get the packet to a specific person in Boston, Massachusetts. If they knew the Boston contact personally, they were supposed to send the packet directly to them. If not, they were supposed to think of someone they did know who was likely to be closer to the person in Boston, sign their name to a roster, and send the packet on to their friend. The friend was then instructed to repeat the process.

Once the Boston contact received the package, he could examine the roster and see how many steps it took for the letter to arrive. While many of the letters were not passed on, 64 of them did reach the final contact person. Among these,

the average number of links from the original recipient to the contact person was between five and six.

Milgram repeated this experiment in different ways, and it has been replicated online more recently (Goel, Muhamad, Watts, 2009). While the number 6 is not necessarily a reliable constant path length between any two participants, one idea clearly emerges from this work: Compared to the number of people in the United States or in the world, the average shortest path between any two is remarkably short.

Small world networks have two primary characteristics: a short average shortest path length and high clustering (measured by the local clustering coefficient). The idea of six degrees of separation reflects this short average path length. Let's look at these attributes more closely, beginning with path length.

"Short" can mean many things. Consider an example: We have a network with 36 nodes and 72 edges. These edges can be distributed in a variety of ways. Figure 3.9 shows what is known as a *regular network*. Each node is connected to a fixed number of neighbors on either side.

There are many steps necessary to find a path from A to B in this network. The shortest path moving clockwise is eight steps. Maintaining this same pattern

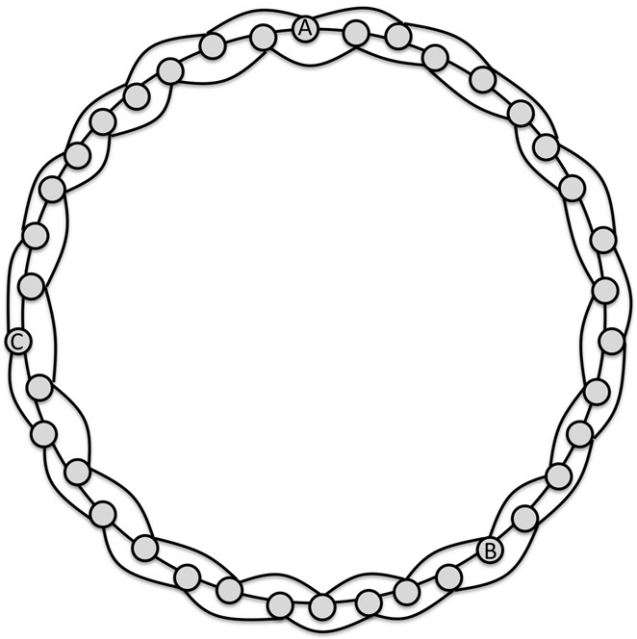


FIGURE 3.9

A regular graph. Each node is connected to the neighbor directly next to it and two steps away in the layout.

of connections and expanding the network to have 1,000 nodes, the average path length would increase by a lot. In Figure 3.9, B is almost halfway around the ring of nodes. Even using the edges that move us two steps around the ring, nearly ¼ of the nodes are touched before reaching B. A quarter of the nodes would still be touched if the graph expanded to 1,000 nodes, so the path length would be around 250. For a node with a million nodes, the path length would be around 250,000.

On the other hand, a graph with the same number of nodes and edges can be created where the edges randomly connect the nodes instead linking them in a regular pattern. This is called a *random graph*, and an example is shown in Figure 3.10.

In Figure 3.10, the shortest path from A to B is much shorter (A to C to B): just two steps. And, the shortest path between most nodes is shorter. The random edges jump from one side of the network to the other, but also connect to nearby nodes. This makes it easy to quickly get near a node, even if it is on the other side of the network, and then reach it through close neighbors.

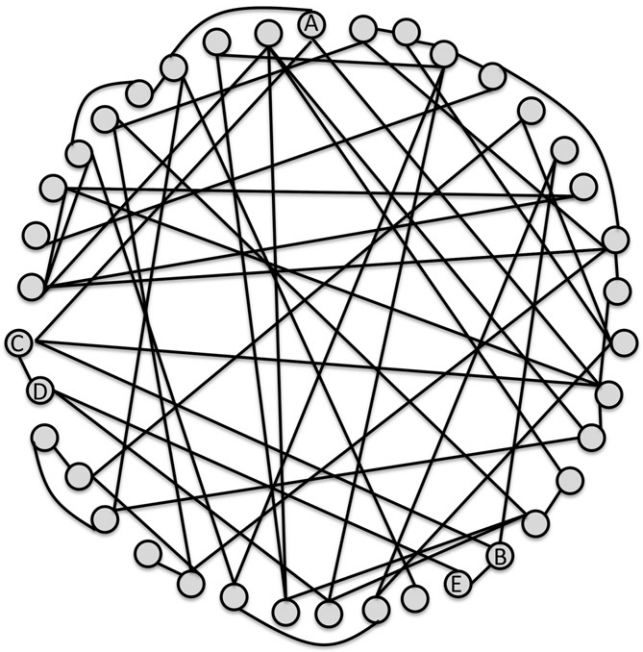


FIGURE 3.10

A random graph, with the same number of nodes and edges as the regular graph shown in Figure 3.9.

If we increased the number of nodes to one million (with a proportional increase in the number of edges), the average shortest path length would increase, but not at the rate we saw in the regular graph. This is because the random edges will still cross the network, making it fast to reach places that would be far away in a regular graph.

Small world networks, including social networks, have this property of a short path length, even when the networks become huge. For example, in late 2011, Facebook studied their network, which had around 720 million users at that point. They found the average shortest path length was 4.74 (Backstrom et al., 2011).

Small world networks have one other main characteristic: high *clustering*. In social terms, this means that a person's friends tend to know one another. Clustering is computed as the average of the nodes' local clustering coefficients.

In Figure 3.9, node A has four neighbors. That means there are six possible edges between them. Three of those edges exist, so the density of A's egocentric network is 0.5. In Figure 3.9, however, A has three neighbors with three possible edges, but only one edge connects them, for a density of 0.33. Node B has the same density as Node A in Figure 3.9—every node has the same pattern of neighbors and connections.

In Figure 3.10, however, none of node B's neighbors are connected, so the density of B's egocentric network is 0. In regular graphs, the clustering is high, but in random graphs the clustering is low.

In 1998, Watts and Strogatz (Watts, Strogatz, 1998) combined these to come up with a way of replicating the structure of small world networks. They took a regular graph and randomly rewired a few edges. These few edges that are moved do not have a significant impact on the clustering, which remains high. However, they are enough to drop the path length dramatically. Even rewiring a small number of the edges—sometimes only 1%—will achieve this. This is illustrated in Figure 3.11.

While social networks and other small world graphs don't usually evolve this way—starting with a regular structure, then gaining a small number of random edges—this work offers interesting insight into how social networks function. The high clustering indicates that many of our friends know one another. In that case, our social connections look a lot like a regular graph. However, we know people in different social circles. We have some connections to different clusters of people (e.g., high school friends, college friends, and co-workers), but also connections to people who may be totally outside our social circle and who connect to none of our other friends. This could be a doctor or dentist, someone we run into at the local sandwich shop, or a friend we met at camp as kids whom we kept in touch with. These people correspond to randomly rewired connections that Watts and Strogatz discussed. In one step, they connect us to social groups that otherwise might be very socially distant.

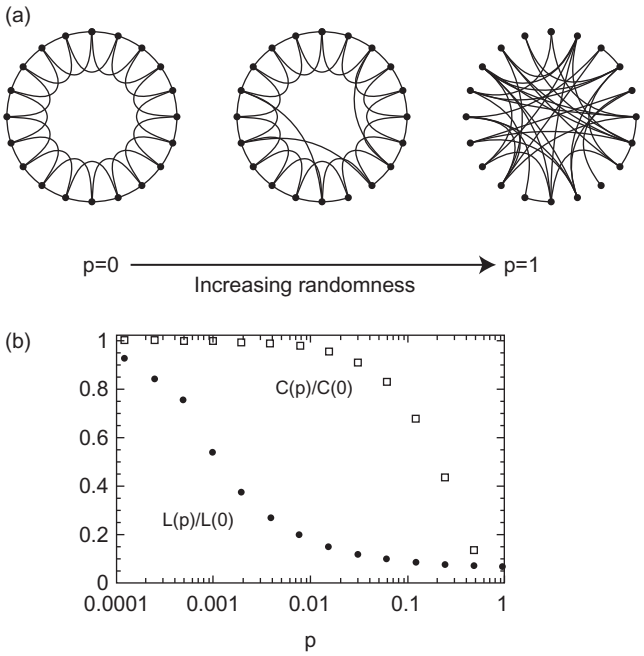
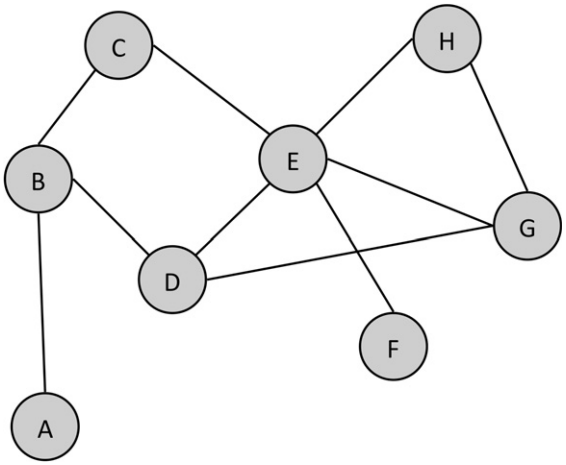


FIGURE 3.11

(a) shows the stages of a regular graph becoming more random by removing and randomly reconnecting some of the edges. (b) shows how the clustering (C) remains high while the average shortest path length (L) quickly drops to low values as the graph becomes more random. The variable *p* indicates the probability of random edge rewiring.

Exercises



1. Answer the following questions about this graph.
   - **a.** What is the degree distribution for this graph?
   - **b.** What is the density of this graph?
   - **c.** Draw the 1.5 egocentric network of node G.
   - **d.** Which node(s) have the highest degree? What is the degree?
   - **e.** Which node(s) have the lowest degree? What is the degree?
   - **f.** Which node has the highest closeness centrality?
   - **g.** Which node has the highest degree centrality?
   - **h.** What is the centralization (based on degree) of the graph?
   - **i.** What is the cohesion of the graph?

DISCUSSION: DENSITY AND NETWORK SIZE

The densities calculated in the examples above are much higher than one would expect to find in most social networks because these networks are small. As the network gets bigger, the density usually goes down. This is because, typically, nodes are connected to only a small part of the overall network. If our network increases from eight nodes to 8 million nodes, it is unlikely that the number of edges will scale up as dramatically. For example, in Figure 3.4, node B is connected to four nodes, but it is unlikely it would be connected to 4 million nodes in a graph of 8 million. Consider a social network like Facebook. If a user has 300 friends but, five years from now, the network has grown to 10 times its current size (not an uncommon feat among social networking websites), we would not expect our user's number of friends to also increase 10 times to 3,000. It might go up, but it will not go up as quickly. Furthermore, even if it did, the density would still decrease. That's because if the number of nodes increases 10 fold, the number of possible edges increases 100-fold.

To see this, recall the formula for obtaining the number of possible edges. If the original network has n nodes, the number of possible edges is $(n^*(n-1))/2 = (n^2 - n)/2$. If the number of nodes increases to 10^*n, the number of possible edges is $(10n^*(10n-1))/2 = (100n^2 - 10n)/2$. Notice the coefficient of n^2 is now 100, not 10.

Plugging in some numbers, suppose the original network had 100 nodes. That would result in $100 \times 99 \div 2$ edges, or 4,950 possible edges. If the network increases 10 times to 1,000 nodes, we have $1{,}000 \times 999 \div 2 = 499{,}500$ possible edges.

Clearly, 499,500 is far more than 10 times greater than 4,950. (Indeed, it is about 100 times as big.) This is because as the number of nodes increases, the number of possible edges increases at that rate squared. So, if the number of nodes increases 10 times, the number of possible edges increases roughly $10^2 = 100$ times. If the number of nodes increases 100 times, the number of possible edges increases $100^2 = 10{,}000$ times.

Another way to think about this is to realize that for a network with n nodes, when one node is added, n possible edges are added. Each of the existing n nodes can connect to this one new node. When two new nodes are added, $n + (n+1) = 2n + 1$ new edges are added. The initial n nodes can connect to the first new node, and then the original n nodes and the one new one can connect to the second new node. As the number of new nodes increases, we add even more new possible edges.

2. Repeat exercise 4 from the previous chapter: List the 15 people you are closest to. Turn this list into a network by listing all the connections between these people. Using this network, answer the following questions:
   - **a.** What is the degree distribution for this graph?
   - **b.** What is the density of this graph?

c. Which node(s) have the highest degree? What is the degree?
d. Which node(s) have the lowest degree? What is the degree?
e. Which node has the highest closeness centrality?
  (i) Which node has the highest degree centrality?
  (ii) What is the centralization (based on degree) of the graph?
  (iii) What is the cohesion of the graph?

CHAPTER

Network Visualization 4

Humans are wired to find patterns visually. We have natural abilities to see anomalies, patterns, clusters, and changes—and we can recognize many of these things without consciously looking for them.

Consider Figure 4.1. Without any instructions on what to look for, and without thinking, you can immediately see the circle in the second row standing out from the pattern of squares.

Similarly, in Figure 4.2, it is easy to see the single outlier point that stands apart from the pattern of values in the chart.

And even in graphs, patterns are easy to see. Consider Figure 4.3.

Even knowing nothing about social network analysis, one can see that node a has many neighbors, there is a tight cluster of nodes in the lower right, and there is a long chain from node b running out to node b4.

In visual data patterns can be recognized that may otherwise be difficult to see in lists of numbers, adjacency lists, or other textual representations of data.

Information visualization deals with the presentation of data in visual format. The data may be numeric, categorical, network data (like social networks), text, and other types. Good information visualization supports users in better understanding the data they are seeing.

The goal of information visualization is to take advantage of humans' natural abilities to see patterns, anomalies, relationships, and features in visual data. Visualization provides an overview of complex data. From there, people can identify features of interest, refocus attention on those features, and explore more. Visualizations are a qualitative way to begin understanding data. From there, quantitative experiments or analysis can follow to explain any insights. Graph visualizations apply all these lessons to looking at the structure of networks.

Future chapters will feature the use of graph visualization for understanding networks in many contexts. This chapter will specifically focus on types of network visualizations and the features that are used to help highlight interesting features.

Graph layout

Every network is made up of nodes and edges. How they are laid out is critical to what an observer is able to understand about a network. There are many types of

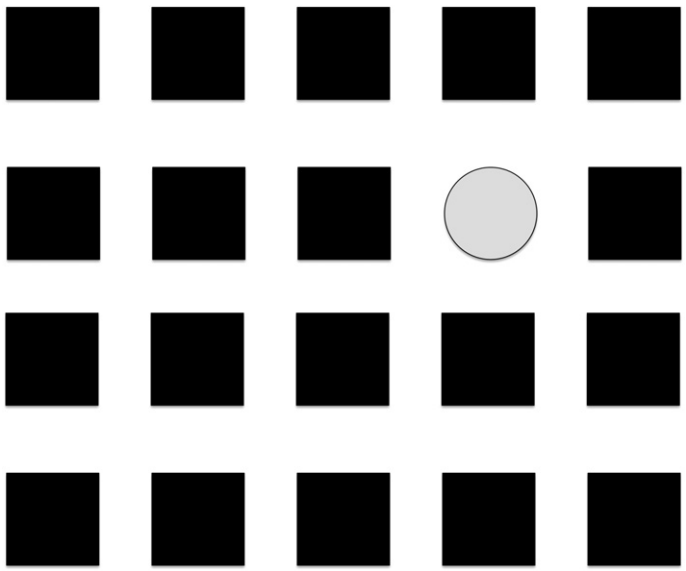


FIGURE 4.1

Without conscious analysis, it is easy to pick out the circle as an anomaly in the pattern.

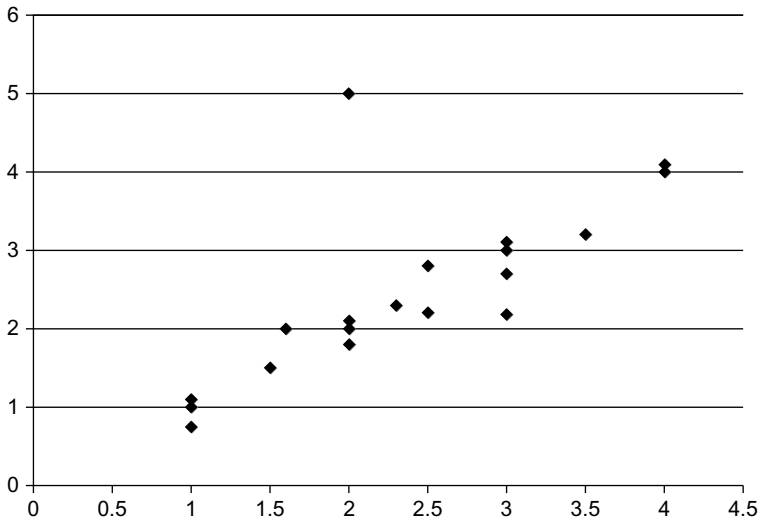


FIGURE 4.2

A single outlier point at value 2 on the x-axis is easy to see separated from the pattern of values.

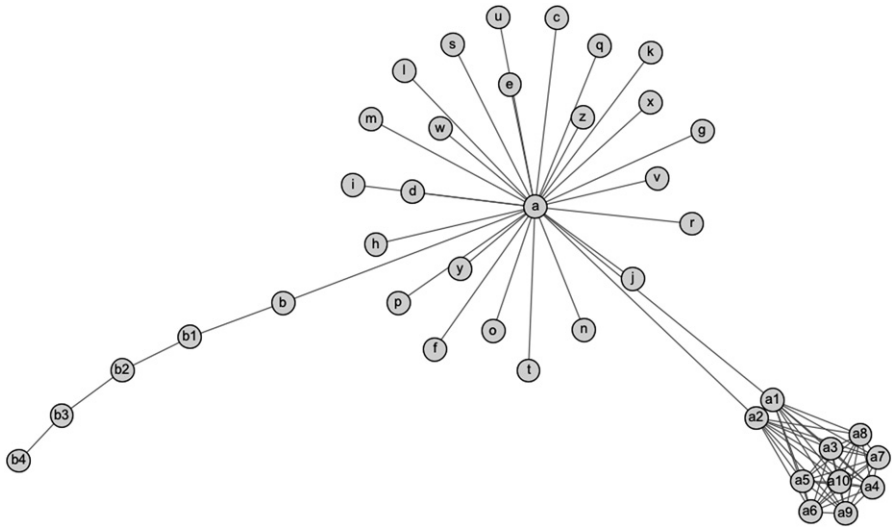


FIGURE 4.3

A sample network visualization.

layout algorithms that position the nodes and edges in different ways when visualizing a network.

What makes a "good" layout is not always clear. It depends on what the analyst wants to find, what type of network is being viewed, and what its features are. However, researchers have presented some general guidelines that make network visualizations easier to work with (Dunne and Shneiderman, 2009):

1. Every node is visible.
2. For every node you can count its degree.
3. For every link you can follow it from source to destination.
4. Clusters and outliers are identifiable.

This section presents a few of the most common network layout algorithms. Note that many of these algorithms have some random features in them. They start with the nodes randomly placed and iteratively move them around into better positions. As a result, running the algorithm multiple times will produce graphs that look different. They will often be similar but may be positioned differently. Also, each iteration helps to improve the layout. Eventually, the iterations make small or no changes. Some applications automatically run the algorithms for a fixed number of rounds, but other times the user can specify a number.

Finally, keep in mind that the absolute position of nodes on the x,y-axis usually has no meaning. There are some graphs that are exceptions, but generally the only thing that matters is how close nodes are to one another. Being positioned to the left or right, top or bottom, does not indicate any properties of the nodes.

Random layout

Often, when loading data into a visualization tool, the nodes are placed randomly. This is called a random layout, and it often does not provide much insight into the structure of the network. Figure 4.4 shows the same network from Figure 4.3 presented in a random layout.

We may be able to tell that node a has a high degree in this network, but the clusters and other patters are not at all clear from the random layout.

Circular layout

Circular layouts place all the nodes in a circle and then add edges between them. Some circular layouts place nodes closer to one another when they are more closely connected. In Figure 4.5, the cluster of nodes labeled a1 through a10 is clearer because of the density of edges in that section of the graph. The chain of nodes from b through b4 is also present, though the edges around the circle are a bit harder to pick up visually than in the Figure 4.3 layout.

A circular layout places nodes in structured positions and then adds edges between connected pairs. Another way to do this is to place nodes in a grid.

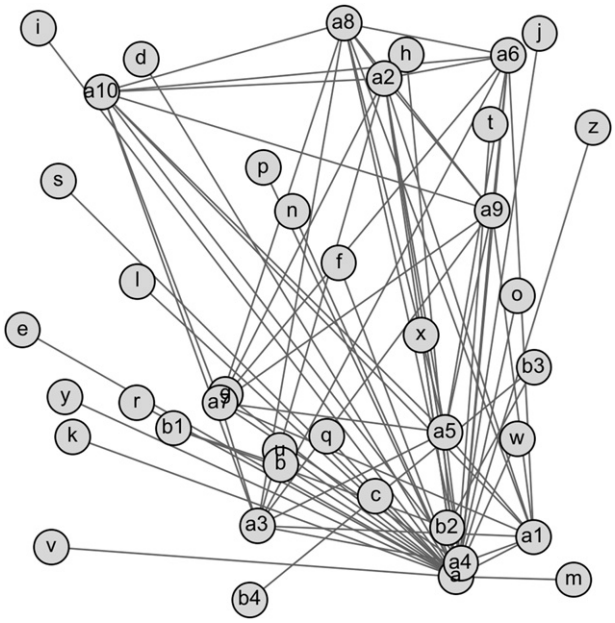


FIGURE 4.4

A random layout of the graph shown in Figure 4.3.

Grid layout

Figure 4.6 shows an example of a grid layout for the same graph in Figures 4.3–4.5. Note that the degree of node a is clearly high, the cluster of nodes a1 through a10 is obvious, and the chain of nodes b through b4 is clear across the top.

Force-directed layout

Most graphs are not laid out randomly or in one of these formats with a predetermined structure. Instead, the layout is dynamic and determined by the connections between the nodes. Those nodes that are more closely connected are laid out close to one another, and those that are distant are shown further apart.

Figure 4.3 uses an algorithm that does this. This type of layout is generally called *force directed*. Nodes and edges are treated as a physical system, and a simulation of that system is applied to determine a final layout. For example, nodes may be treated as objects, and edges may be treated as springs that apply equal force. The nodes are randomly laid out, connected by springs for edges, and then a simulation of how the springs would physically behave determines the final position of nodes and edges. A cluster of nodes with many connections will be

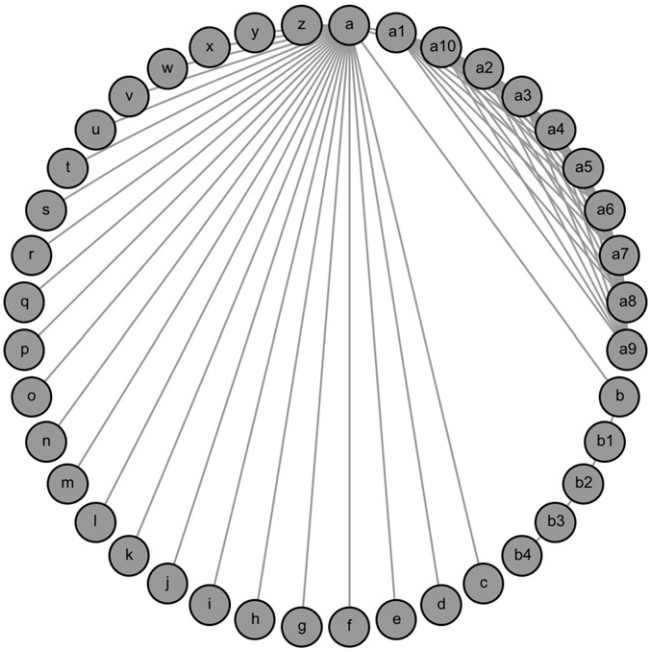


FIGURE 4.5

A circular graph layout for the same graph shown in Figure 4.3.

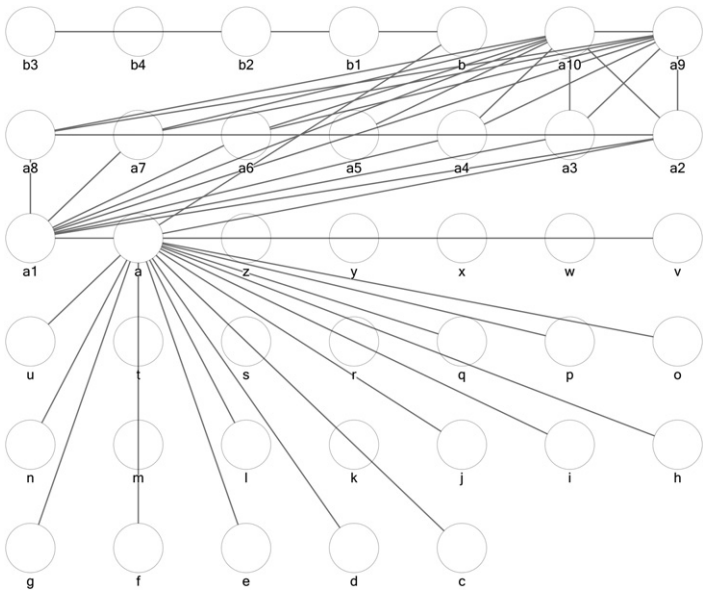


FIGURE 4.6

A grid layout of the modes in the sample graph.

close together, because pulling any node away pulls on many springs that want to keep it close. Nodes with little or no connection are not attracted to one another. Similar approaches to layout that rely on physical simulation include simulated annealing or treating the nodes like charged particles.

Yifan Hu layout

Many algorithms lay out graphs in this matter. Figure 4.3 uses one called Yifan Hu. Figure 4.7 uses a variant called Force Atlas. While there are differences between Figures 4.3 and 4.7, the similarities in clustering and separate nodes are clear.

Harel-Koren fast multiscale layout

The Harel-Koren fast multiscale algorithm (Harel and Koren, 2000), available in NodeXL, is designed to quickly lay out large, complex graphs. It is based on force-directed layout algorithms but uses optimizations in the underlying code to make the algorithm computationally efficient. For large graphs with thousands of nodes, generating a layout with many force-directed algorithms can take a very long time. With Harel-Koren, it often can be achieved in a few seconds, making it an ideal choice for large networks.

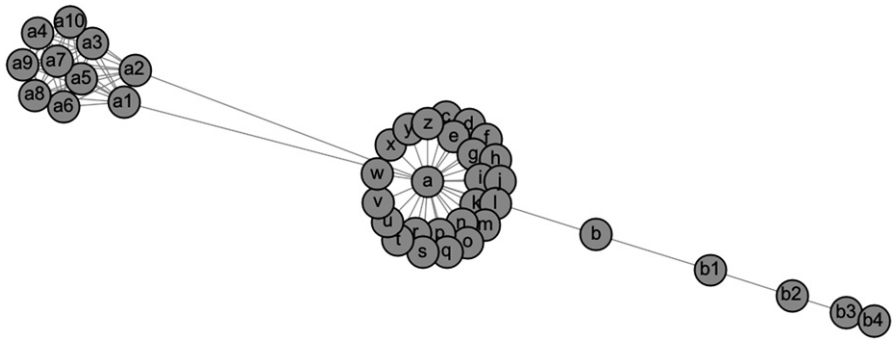


FIGURE 4.7

A layout of the sample network using the Force Atlas algorithm.

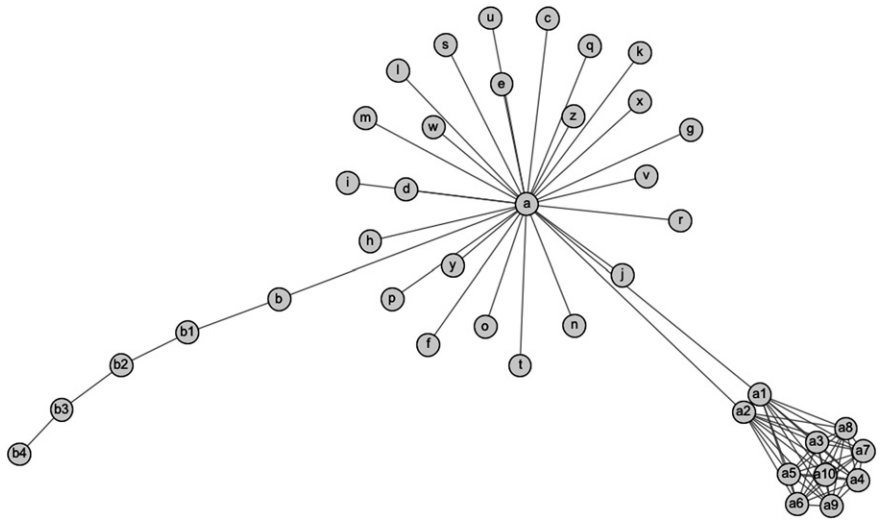


FIGURE 4.8

The graph laid out with the Harel-Koren Fast multiscale algorithm.

Other layouts

Most graphs will be presented using a force-directed layout algorithm. However, there are some more sophisticated layouts designed to convey additional information through layout. Figure 4.9 shows a layout available in the graphing program NodeXL. Here, nodes are clustered, grouped into boxes, and then links are added within and between boxes.

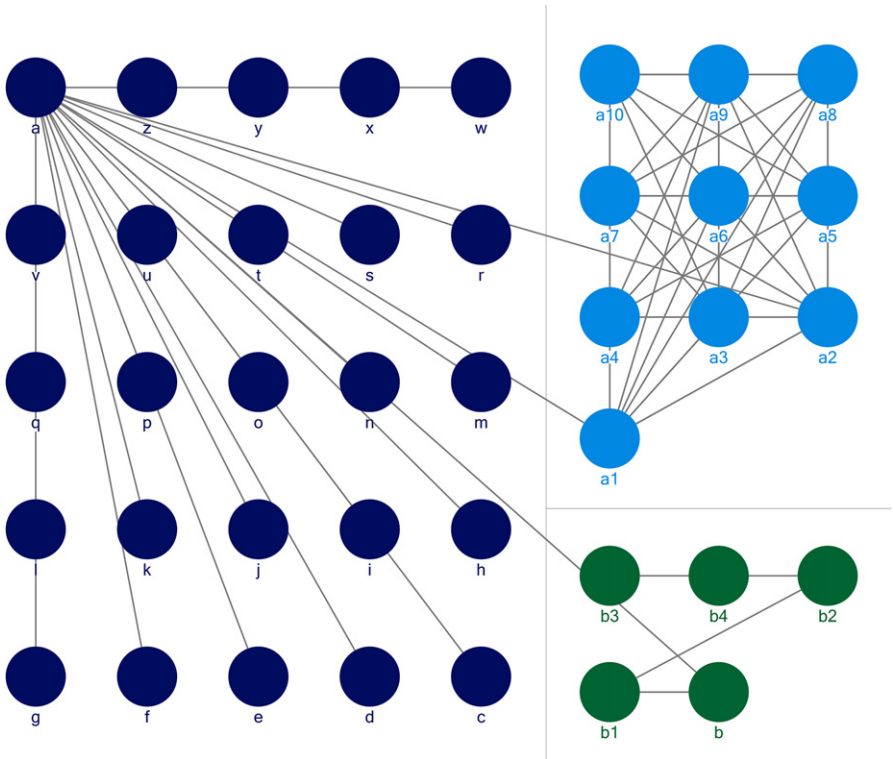


FIGURE 4.9

A layout that groups clusters into boxes, sized by the size of the cluster, and shows links between boxes.

Network visualization is an active area of research, so that new and creative mechanisms for visualization are constantly being developed. The examples above show the most commonly used and core methods of visualization, but network analysis tools will likely have additional options designed to support new and interesting types of analysis.

Visualizing network features

The layout algorithms discussed in the previous section dictate the placement of nodes and edges. Other network features, like edge weights, node properties, labels, and clusters, can also be visualized. Like the layout algorithms, there are many options to do this. This section will present some of the most common ways this is done.

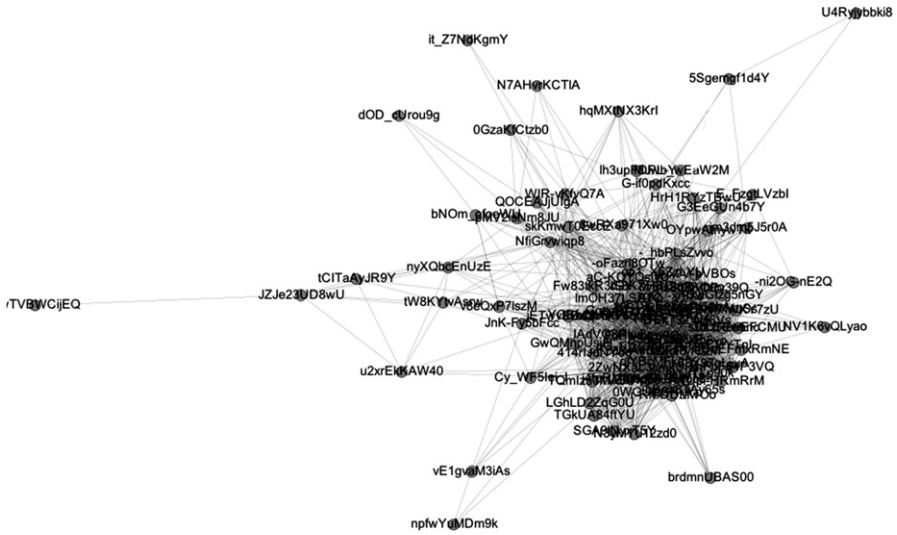


FIGURE 4.10

A network of YouTube videos with the node labels shown.

Labels

Labels are some of the more difficult attributes to show in a network, both on nodes and on edges. The example graphs in the previous section all have node labels, but the graphs are small and the labels are short. Figure 4.10 shows a network with only 92 nodes, which is still relatively small. The nodes represent YouTube videos, and the edges indicate that they were tagged with at least one similar term. The node labels are the YouTube identifiers for each video. Even in this small graph, the image becomes very cluttered with all the labels shown.

Similar problems happen with edge labels. Whether shown on top of the edge with straight alignment or angled along the edge, the graph tends to become cluttered and difficult to read. Some techniques can improve on this a bit, either by putting boxes around the text, by only showing a few labels of interest, or by relying on interactive interfaces that only show labels on demand. The latter allow the user to move the mouse over a node or edge and see the label or other data on demand. That facilitates exploration of the graph without the clutter. Still, there are no solutions to totally eliminate this problem when producing fixed visualization images, so often labels are left off.

Size, shape, and color

Fortunately, showing other attributes of nodes and edges in graphs can be easier. Categorical or quantitative attributes are particularly easy to show by adjustments

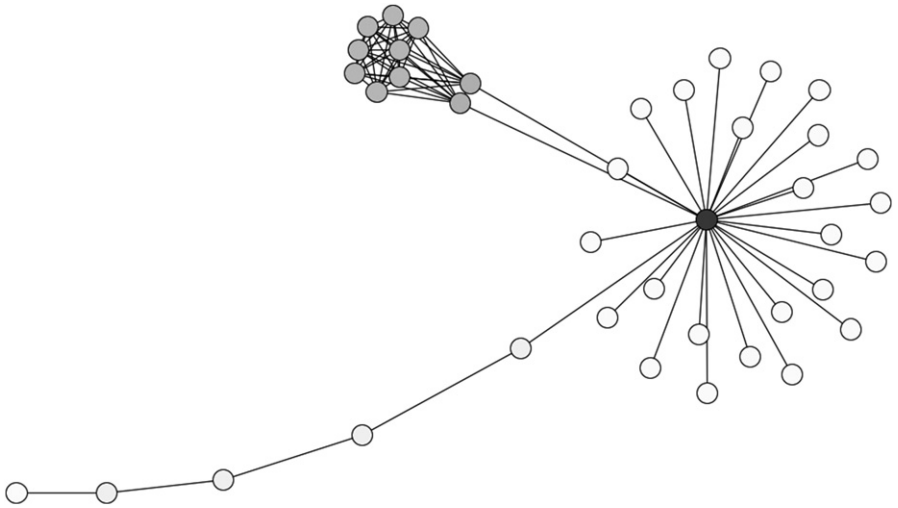


FIGURE 4.11

Color-coding nodes according to their degree, with higher degree shown by darker nodes.

in size, shape, or color. Return to the example graph used in Figures 4.3–4.8. There are many statistics about the nodes in that network: degree, centrality, and so on. These can be encoded using color, size, or both. Figure 4.11 shows color encoding of node degree. Darker colors indicate nodes with higher degrees, and not surprisingly, node a is the darkest. For clarity, the node labels have been left off this graph.

Node color could also be used to indicate other attributes of a node. For example, in a visualization of a person's email network, node color could indicate if each person is a friend, family member, classmate, co-worker, and so forth.

Keeping color as an indicator of degree, node size can be used to indicate other attributes. For example, clustering coefficient is interesting here, since there is a tight cluster where all the nodes are connected, while in the rest of the graph, the clustering coefficient is very low for each node. Figure 4.12 shows a graph that uses color for degree and size for clustering coefficient.

Edges can also be treated with color or thickness to indicate their attributes. For example, different types of relationships could each be coded in a different color. Edge weights are also commonly visualized. These could indicate the strength of a relationship, the frequency of communication, or other factors. Figure 4.13 shows the same example network with weights added to the edges. These are visualized by adjusting the width of the edge. Wider edges indicate stronger relationships.

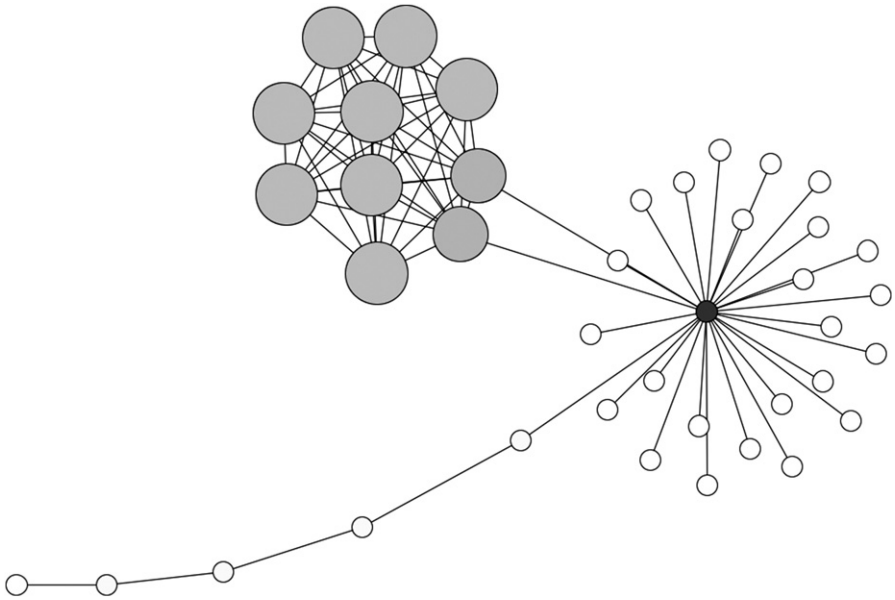


FIGURE 4.12

A graph indicating clustering coefficient with node size and degree with node color.

Larger graph properties

Larger graph properties can also be encoded in visualizations. For example, clusters are sometimes apparent on their own (like the group to the upper right in Figure 4.11), but visual properties to indicate them will often clarify a visualization further. Figure 4.13 shows a new graph that has two main clusters. This graph is a network of YouTube videos, where nodes represent videos and edges connected videos that share a common tag. All of these videos were tagged with the word "cubs"; this example will be discussed more in Chapter 7. Even without the color coding, the two groups would be relatively easy to see. But using a community detection algorithm that groups nodes into clusters, and then color coding by those clusters, makes it even more apparent. This is shown in Figure 4.14.

Scale issues

The example networks shown so far have been relatively small—a few hundred nodes and a few thousand edges. Visualization is very useful for analyzing networks of this size or smaller. When networks become much larger, the quality of the visualization diminishes.

Figure 4.15 shows a network from a peer-to-peer file sharing network. Nodes represent hosts (computers participating in the network), and edges represent

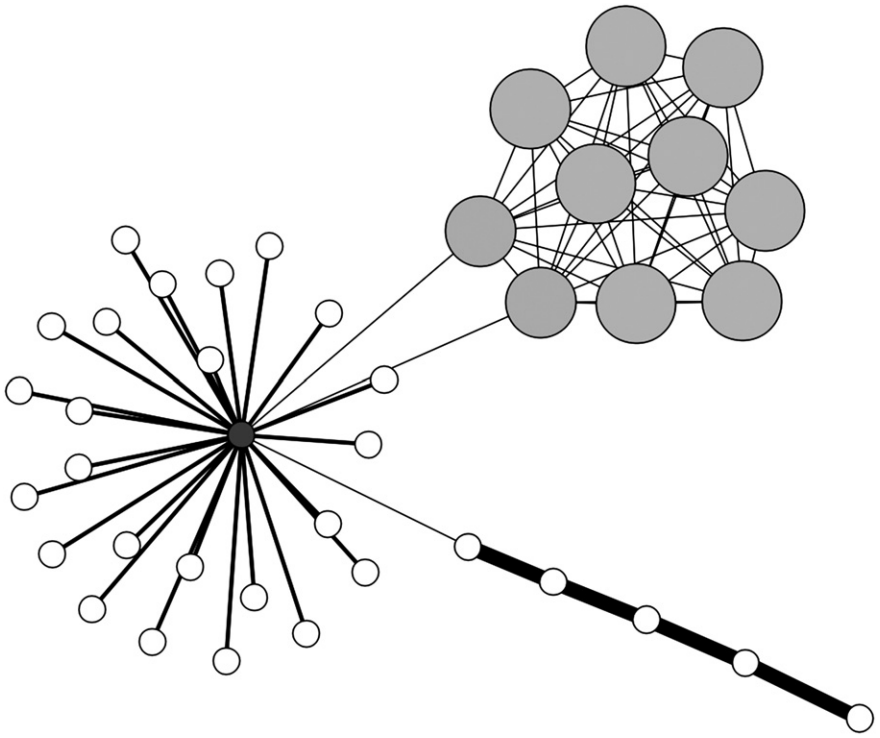


FIGURE 4.13

The sample network with edge width indicating the weight on each edge. Note that the central node has medium-strength relationships with most neighbors, but weak ones to the cluster in the upper right and the chain in the lower right. The chain of nodes in the lower right have high weights on the edges connecting them.

connections between them (usually one computer downloading a file from another). There are close to 11,000 nodes in this network with roughly 40,000 edges. Even with a very low density (<0.001), there are still too many nodes and edges to see much of anything.

Depending on the structure of the network, it is sometimes possible to get useful visualizations with up to around 10,000 nodes; however, networks under 1,000 nodes are typically safest.

Density

Density can also be a problem for visualization, even if the number of nodes is small. Figure 4.16 shows a network of members of the U.S. Senate. There are only 100 nodes but over 4,100 edges. The edges indicate that the senators have

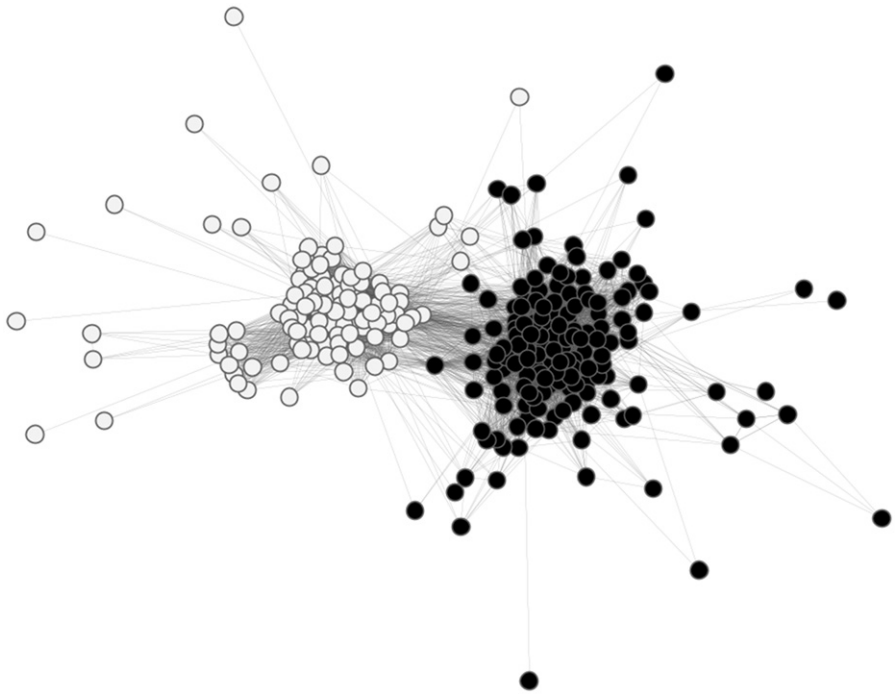


FIGURE 4.14

A network of YouTube videos where color indicates the community or cluster to which each node belongs.

voted the same way in at least one bill. The edges have a weight, indicating the percentage of bills on which the two senators have voted in the same way. Figure 4.13 has the edges filtered so that only those with a weight of 40% or more are visible. However, as this network shows, there are no interesting patterns visible with the threshold of 40%; the network is simply too dense.

Filtering for visual patterns

It is often difficult to see any patterns in very dense networks. One way to compensate for this is to filter the networks when possible. For example, if we take the same network from Figure 4.13 and filter the edges so that they only connect senators who have voted the same way on at least two-thirds of the bills, the pattern changes dramatically. This is shown in Figure 4.17.

In this figure, two clear clusters emerge, representing the two major political parties. Furthermore, five senators are pulled out from the major party clusters along the center, indicating that they frequently vote with members of both parties.

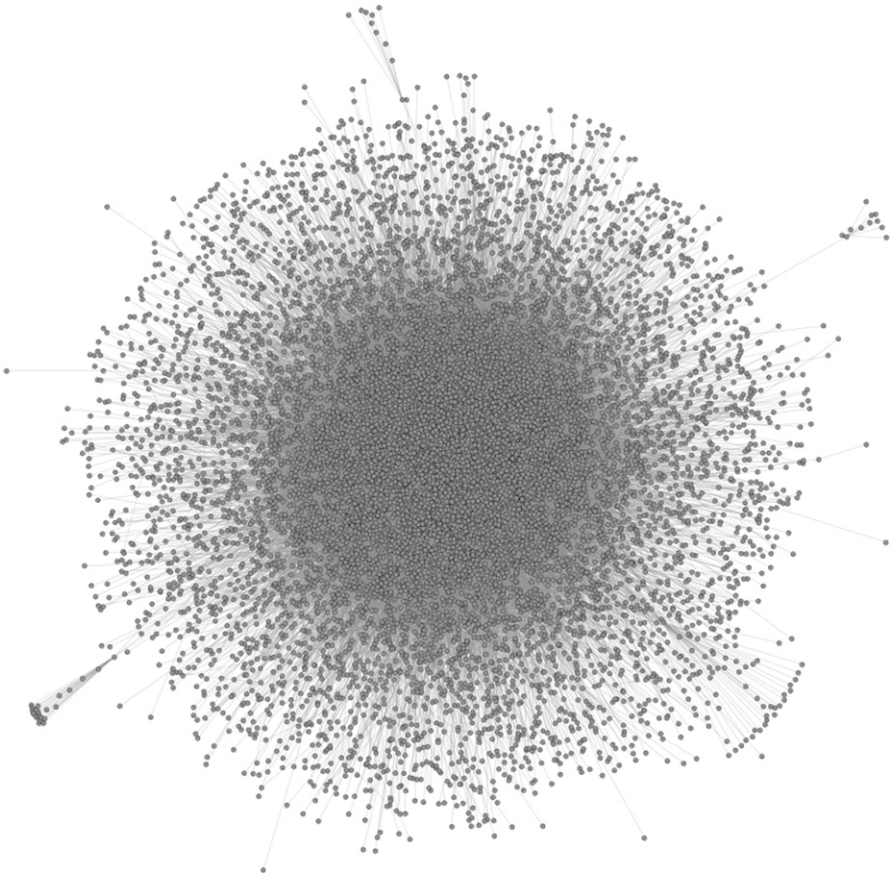


FIGURE 4.15

A network with 11,000 nodes and 40,000 edges.

Graph simplification

An active area of research in network visualization is graph simplification. Because large networks are very common when working with social media, problems of scale are common. Graph simplification techniques include grouping clusters of nodes into a single node and representing the edges between clusters as a single edge, representing structural patterns as representative shapes, or showing only part of the graph at a time. As an example, Figure 4.18 shows a tree-structured network visualized with a tool called Space Tree (Plaisant et al., 2002). The nodes and edges to the right of the first level are hidden but are summarized with triangles. The size, color, and angle of the triangles indicate the depth, number of nodes, and width of the summarized structure.

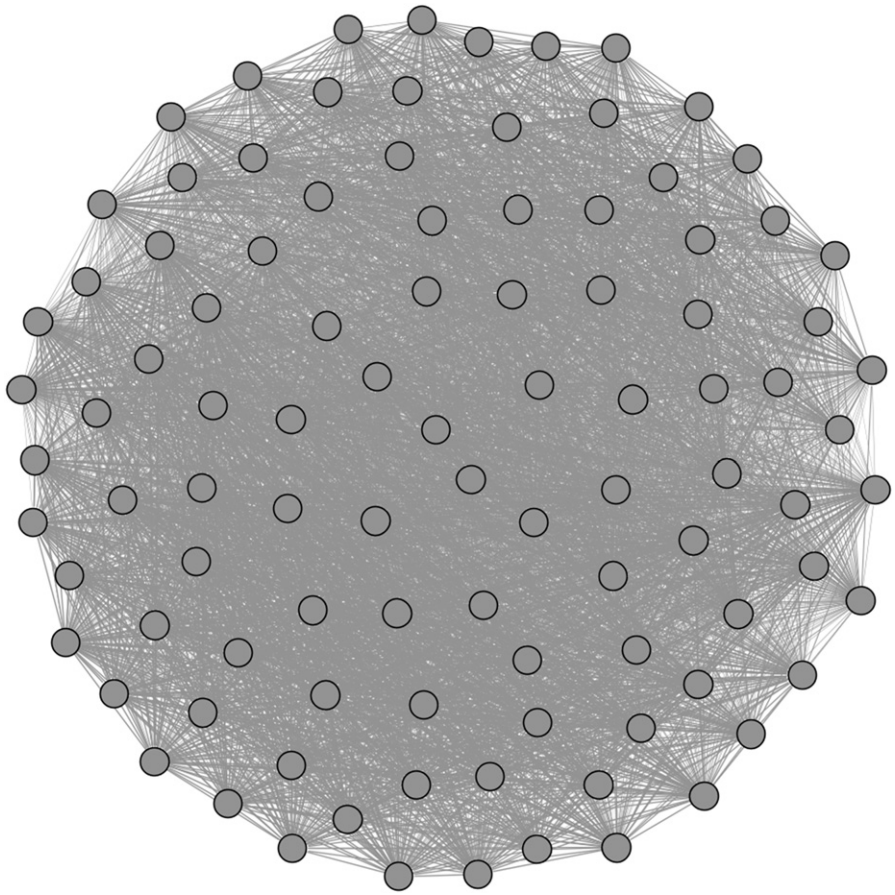


FIGURE 4.16

A network of senators (nodes) with edges connecting senators who have voted the same way at least 40% of the time. The network is very dense, so it is not possible to see any interesting patterns.

Figure 4.19 shows a technique called *motif simplification* applied to a network built from the Lostpedia website. The graph on the left is summarized in the graph on the right with arcs to represent nodes that have many singleton neighbors (called fans), and the many nodes that are linked to two of the discussion topics are summarized with arches.

Many methods are availablefor simplifying graphs, and these may make it possible to find patterns in large or dense networks when traditional visualization methods would fail.

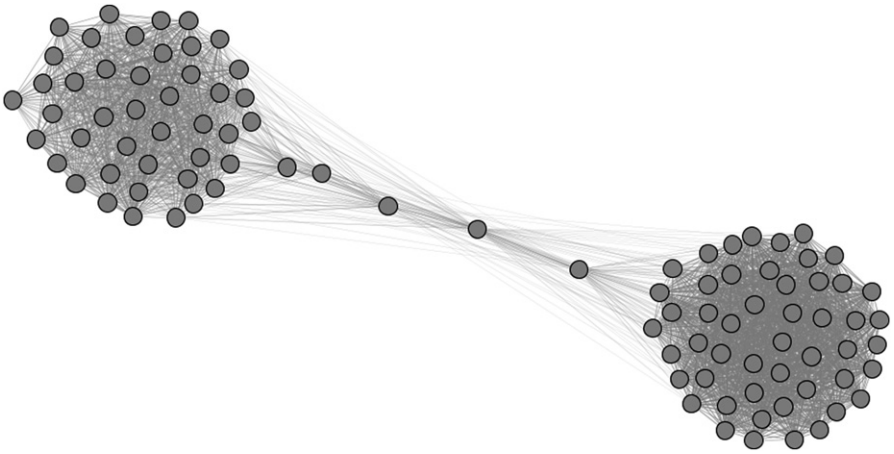


FIGURE 4.17

The same network of senators as shown in Figure 4.13, now filtered to include only edges between senators who have voted the same way on at least two-thirds of bills.

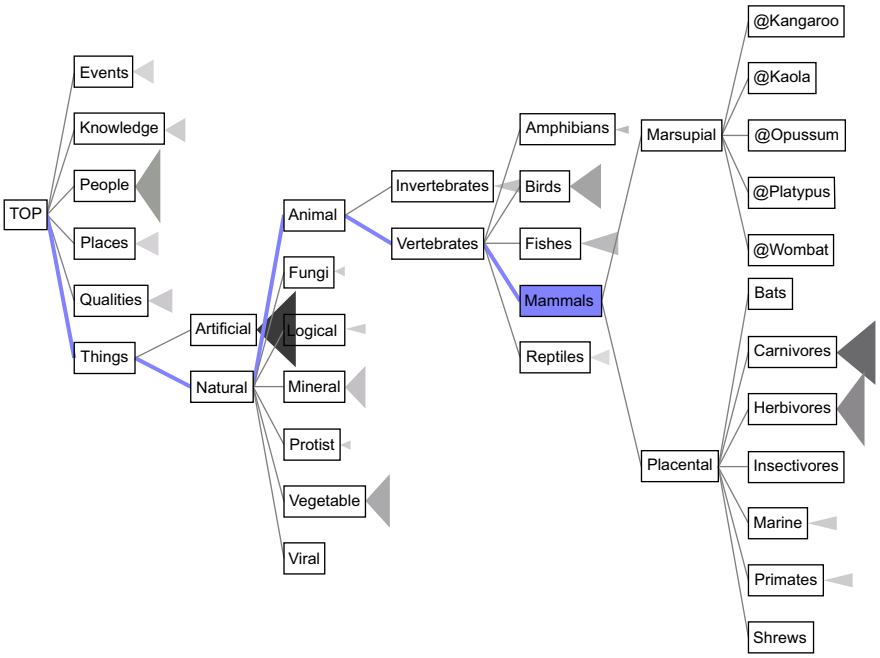


FIGURE 4.18

A tree-structured graph that uses triangles to summarize the nodes and edges that follow after a given node.

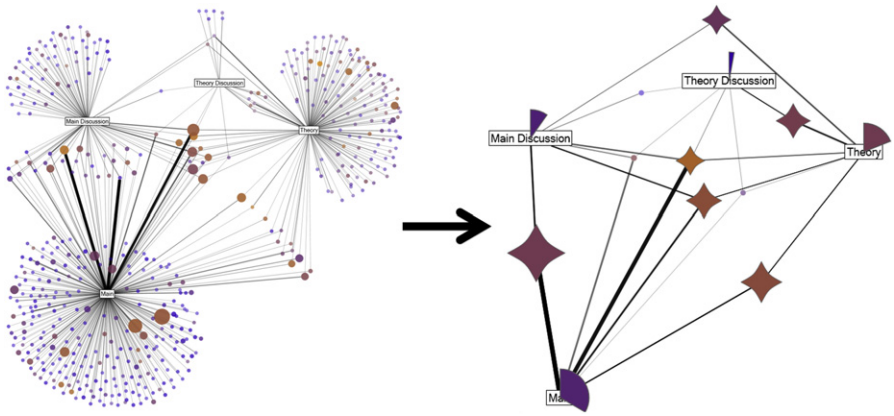
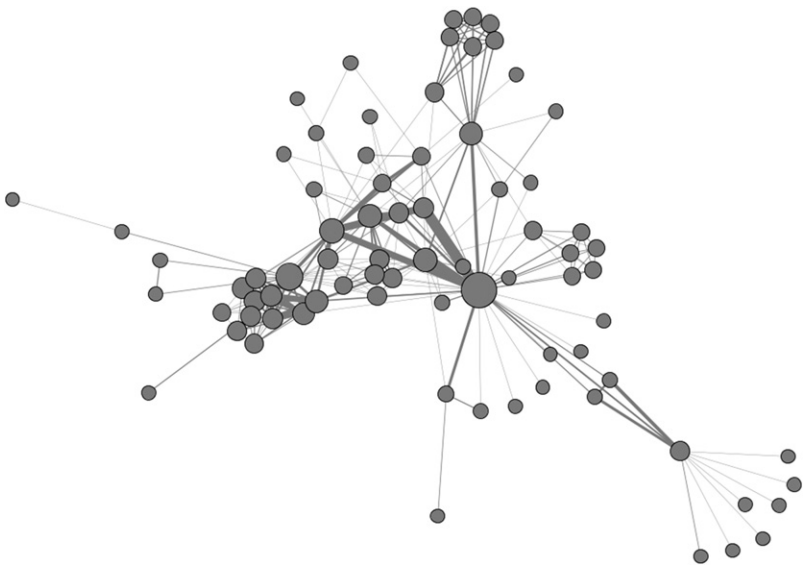


FIGURE 4.19

The graph on the left is summarized using simple glyphs into the graph on the right. This uses a technique called *motif simplification* (Dunne and Shneiderman, 2012).

Exercises

1. The graph below is a network of characters in the Victor Hugo novel *Les Miserables*. Larger nodes have a higher degree, and thicker edges indicate stronger relationships.



   a. Which node or nodes seem most important in this network? Which seem least important? What in the visualization leads you to draw your conclusions?
   b. List two other interesting structural features you see in this graph. These could be hubs, clusters, nodes with unusual structural properties, or any other interesting feature. Describe what insights you learned from these structural properties.
   c. Download the dataset for this network from the book's website (or elsewhere on the Internet) and visualize it in your favorite graphing program (e.g., Gephi or NodeXL). Use the software's features to explore the names of the nodes and identify who is in each of the interesting structural features you listed in (b).
   d. Using the visualization software and network from (c), add other features to the network that offer new insights. This could be adjusting the color of the nodes to indicate a certain feature, clustering nodes, or showing labels. Describe your new visualization and how the features you added help understand the network.
2. Create a list of 10 of your Facebook friends and make an adjacency list of the connections between them. Visualize this network using Gephi or NodeXL.
   a. Which friend has the highest degree?
   b. Which friend has the highest centrality?
   c. Are there any obvious clusters?
   d. Are there any nodes that are outliers?
3. Using the *Les Miserables* dataset above, try at least three different network layout algorithms to visualize the graph. Show all three visualizations, describe each, and explain which visualization you think is better and why.
4. Using a spigot in Gephi or NodeXL, import an existing network. This can be a Twitter user's network, your personal email network, a network of YouTube videos, or the like.
   a. Create two visualizations of the network that tell different stories. Use color, shape, size, and layout to indicate interesting features of the network.
   b. Write a paragraph explaining each visualization.
   c. Write a paragraph comparing the two visualizations.
5. Do a web search for graph visualization examples (an image search will produce many examples). Choose one good one and one bad one. Explain the good and bad points of each. Analyze each according to the four principles of good graph layout presented early in the chapter.

CHAPTER

Tie Strength 5

Social relationships are complicated. The type of relationship people have will draw on many things like their history and similarity, each person's personal background and preferences, environmental factors, and more. Relationships are also multifaceted, and many relationship types can be used in social network analysis. One of the most useful is the idea of tie strength.

Tie strength is a measure of the strength of a relationship between people. The concept was introduced by Mark Granovetter in 1973. He asserted that "the strength of a tie is a . . . combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (Granovetter, 1973).

While there is a range of tie strength, Granovetter defined two main types: *strong ties* and *weak ties*. Strong ties are rare and are usually family members or very close friends. These are usually people a person sees frequently, with whom one shares personal details of one's life, and for whom the person will do and expect favors. Weak ties are much more common and include acquaintances and more casual friendships. Of course, there is a spectrum of tie strength, and any relationship may fall along the scale from weak to strong.

People with whom someone has no meaningful relationship—the familiar stranger one passes on the street and nods to, or a vendor that a businessperson may contact—are not considered in the spectrum of strong or weak since there is not much of any relationship present. These are sometimes called *absent ties* and would not appear as edges in the social network.

Tie strength is a very important factor to consider in social network analysis. Consider the flow of information through a network. Weak ties often connect to diverse groups of people with different perspectives. These ties allow information to move throughout the network. Strong ties are more trusted, and their information is more likely to be reliable. The same features apply when considering the spread of other things through a network, like a disease. Someone is more likely to catch a cold from a weak tie (because there are many of them, and they will carry germs from many different groups of people). But because of the high level of close contact, it will likely spread quickly to one's strongest connections.

That is not to say that tie strength is the only factor influencing trust, reliability, and closeness in social networks. Weak ties may provide highly trusted information; for example, a physician may be more trusted about medical information than someone's family members. In this example, the authority of the physician

outweighs tie strength. Throughout the chapter, we will discuss factors that influence tie strength and its role, but there will be exceptions to every example. Thus, tie strength cannot be treated as the only factor influencing relationships, and observing people's interactions cannot predict it perfectly, but the guidelines presented here are useful for considering this important relationship measure and its role in networks.

This chapter will address how to measure tie strength, see how it relates to network structure, and learn how it affects the way information, diseases, and more spread through social networks.

The role of tie strength

One of the first efforts to understand the importance of tie strength was Granovetter's study on how people get jobs. He published the results of the study in his paper, "The Strength of Weak Ties" (1973), and again in greater depth in his book *Getting a Job* (1974). His research studied men in a Boston suburb. Through interviews and surveys, he looked at how they found or received information when they changed jobs. Granovetter reports many scenarios like the following:

> *Carl Y. was doing commission sales for an encyclopedia firm, but was not doing well. He decided he would have to find a different job; meanwhile, he started driving a cab to bring in extra money. One passenger asked to be taken to the train station where he had to meet a friend. This friend turned out to be an old friend of Carl Y's, and asked him "what're you doing driving a cab?" When Mr. Y explained, the friend offered him the job he now holds—labor relations manager for a small company, owned by his friend.*
>
> **Granovetter, 1974, p. 34**

> *George C. was working as a technician for an electrical firm, with a salary of about $8000, and little apparent chance for advancement. While courting his future wife, he met her downstairs neighbor, the manager of a candy shop, a concession leased from a national chain. After they were married, Mr. C. continued to see him when visiting his mother-in-law. The neighbor finally talked him into entering a trainee program for the chain, and arranged an interview for him. Within three years, Mr. C was earning nearly $30,000 in this business.*
>
> **Granovetter, 1974, p. 49**

> *Edward A., during high school, went to a party given by a girl he knew. There, he met her older sister's boyfriend, who was ten years older than himself. Three years later, when he had just gotten out of the service, he ran into him in a local hangout. In conversation, the boyfriend mentioned to Mr. A. that his company had an opening for a draftsman. Mr. A. applied for this job and was hired.*
>
> **Granovetter, 1974, p. 76**

In all of these examples, the people obtained their jobs through social contacts rather than responding to an ad, applying directly, or going through a recruiting service. Indeed, Granovetter found that 56% of the people he talked to found their jobs through personal contacts (Granovetter, 1974, p. 14).

Furthermore, all of these examples illustrate people getting information about jobs from relatively casual relationships. None of these are cases where someone was hired by their life-long friend or a close family member. Every case has at least one weak relationship link—an old friend, a neighbor of the family, or someone met at a party. These relationships, known as weak ties, turn out to be incredibly important for finding a job and for the spread of anything through social connections. This is because a person's weak ties (acquaintances) are likely to travel in different social circles, while strong ties are likely to know one another and travel in the same social circles. The social connections of weak ties are more diverse and provide access to a much broader range of information and people than strong ties. They are also much more plentiful than strong ties.

This was emphasized in a replication of Milgram's "six degrees" experiment. Researchers gave booklets to participants and instructed them to pass the booklets on until they reached an unknown target person. At each step the participants recorded to whom they gave the booklet and how they knew that person. Results showed that chains where the booklet successfully reached the target made much heavier use of weak ties (Lin et al., 1978).

Weak ties are important in ways beyond the spread of information. They also play an important role in how organizations and groups function. A decade after his original article was published, Granovetter (1983) presented two diverse examples of how weak ties help in integrating social groups.

In one study, researchers found that one way to improve racial integration in classrooms was to arrange class structures to form many weak ties between black and white students, rather than focusing on building fewer, stronger relationships between students of different races (Karweit, 1979). Another study looked at job satisfaction in a children's psychiatric hospital. While many hospitals of this type have high turnover rates and low job satisfaction, this particular hospital was quite different. Morale was quite high, and the researcher attributed this to the many weak ties among the hospital staff. Instead of being organized into tight-knit, insular groups, the staff all frequently interacted with one another and everyone knew each other on a first-name basis. This made it easier for individuals to interact with one another and to integrate into new groups (Blau, 1995).

This does not mean that strong ties are unimportant. Strong ties—family and close friendship relationships—are more committed, reliable, and trustworthy. They also form a critical part of social structure. Even in a job search, where Granovetter originally showed the importance of weak ties, strong ties play an important role. A follow-up study of nearly 1,800 people living in Philadelphia showed that over 56% of them used ties to find their job, but among those, 72% used *strong* ties.

Strong ties are more willing to help and have greater motivation to do so. This is true in finding jobs, but in many other contexts as well. Two simultaneous studies of social structure in poor neighborhoods in the U.S. Midwest (Stack, 1975) and in Mexico City (Lomnitz, 1977), found that people rely heavily on strong ties. People with strong ties provide one another with access to food stamps, housing, child care, money, social support, and other items of value. The close-knit relationships of both family and family-like friends are vital to this social structure. Since resources are so scarce, sharing among strong ties keeps the people in the communities going.

Both types of ties have their benefits. Because people linked by strong ties see one another frequently and interact on a deep level, they are motivated to help one another, put effort into the relationship, and behave in a trustworthy way. Weak ties, on the other hand, do not have these motivations, but they do provide access to a more diverse set of information and resources. They are also easier connections to form and maintain. This means they provide an easy and important link to the world outside a person's core social circle.

Measuring tie strength

To analyze tie strength in social network analysis, the network must include relationship information. In small networks, especially if data is hand-collected, it may be feasible to ask each person to rate the strength of their tie to each person. By necessity, larger networks require a mechanism for measuring tie strength. There is no single factor that defines a strong or weak tie, but a number of predictors can be combined to estimate the strength of a relationship.

In his original paper, "The Strength of Weak Ties," Granovetter offers four intuitive factors that may contribute to tie strength. As stated above, he writes, "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."[1]

Time can include the amount of time people spend with each other, the duration of their relationship (i.e., how long they have known each other), and how frequently they see one another.

Emotional intensity is indicated by the closeness of a relationship; close friends or family members are likely to be strong ties, while more casual friends and acquaintances would be weaker ties.

Intimacy, or mutual confiding, relates to people sharing secrets or intimate personal details with one another. The more of this information they exchange, the closer their relationship is likely to be.

[1] p 1361.

Reciprocal Services are favors that people do for one another. They may be personal (e.g., pet sitting or picking up someone's dry cleaning), financial (e.g., loaning money), professional (e.g., putting people in contact with one another), or otherwise.

Since originally proposed in 1973, researchers have investigated what other factors might also play a role in tie strength. There are several of these factors, but three are more widely accepted as important.

Structural features relate to the social network of the two people in question. Those who have many mutual friends are likely to have stronger ties.
Social Distance measures how different people's social situations are. This includes factors like age difference, race, education, and socioeconomic status. People with strong ties tend to have similar social attributes.
Emotional Support describes the communication between people that validates their emotions, shows understanding of their problems, and tries to alleviate stress.

These seven factors are not equally important in determining tie strength (although there is not total agreement about their relative importance). For example, studies have consistently shown that measures of a relationship's closeness, often captured through emotional intensity or intimacy, are among the strongest indicators of strength (Marsden, 1984).

These factors are not independent. For example, people who have a very intimate relationship will often spend a lot of time together. People of different ages and positions in life, or those who have a large social distance are also less likely to have as many mutual friends as people with similar social positions. Thus, when measuring behavior or interactions, a single measurement may describe more than one of these factors.

Additionally, it does not always follow that having many of these factors indicates a strong tie. For example, roommates may have many friends in common, be in socially similar situations (and therefore have a low social distance), spend a lot of time together, and even do favors for one another, yet still maintain a distant and impersonal relationship.

A natural question to follow is how these factors are measured. Intimacy, for example, is difficult to quantify, and depending on the context of a relationship, its meaning may vary. Indeed, there is no single correct answer for how to measure any of these relationship features. If measuring them is important, it will depend on the context, the information available, and likely many other factors.

An interesting example of one way this measurement has been done is presented in Gilbert and Karahalios's work on computing tie strength in social media (2009). In their study, subjects answered a series of questions about their relationship with friends on Facebook, and information was collected from both users' profiles and their interactions. This Facebook data was used to create a set of attributes designed to reflect each of the seven aspects of tie strength mentioned above.

Here are just a few examples of the over 70 variables they used to measure tie strength in their study.

Intimacy
- Number of days since their last communication
- Number of friends in common
- Number of "intimate" words in their communications, as determined by software that automatically analyzes text

Intensity
- Number of words exchanged on one another's walls
- Depth of email threads in their inboxes (i.e., how many messages were sent back and forth in a conversation)

Reciprocal services
- Number of links shared on one another's wall
- Applications the users had in common (presumably because they could be working together within the application context)

Social distance
- Age difference
- Difference in the number of educational degrees
- Difference in the number of occupations

Together, these variables were used to try to predict the tie strength of two people's relationship. They worked quite well—the researchers showed that when users rated their tie strength on a scale, the automated method could predict it to within around 10% of its actual value.

While Facebook offers a large dataset that can be measured for each aspect of tie strength, Facebook-specific features will not be available in most contexts. Determining tie strength by analyzing email communication, for example, will differ from tie strength being inferred from a series of blog posts or by studying interactions on a discussion forum. The seven factors above can serve as a guideline, but ultimately it is an understanding of the relationships and interactions in a given context that will dictate how it can best be measured.

Tie strength and network structure

We learned above that network structure is related to tie strength. People who have many friends in common are likely to have stronger ties than people with few mutual friends. There are many ways that tie strength and network structure are related. This section will consider several of the most significant relationships.

Strong ties have unique properties within a social network. They are not randomly scattered throughout the network, but rather tend to appear in clusters. As an exercise, think of five to seven of your strongest relationships. Write these in a

circle, and draw connections between the people who have relationships with one another. Use thicker lines for the strong ties between these people. Very often, there will be many strong ties among the people with whom you share strong ties.

This illustrates the tendency of strong ties to appear in clusters. Each person will have many more weak ties connecting them to people outside this small group, but a person's strong ties tend to have strong ties to one another.

This pattern of strong ties being densely connected leads to another structural concept called the *forbidden triad* (see Figure 5.2). Imagine three people: Alice, Bob, and Chuck. Alice and Bob have a strong tie, and Alice and Chuck also have a strong tie. What does that tell us about the relationship with Bob and Chuck? While we cannot draw any absolute conclusions, it is likely that some sort of tie exists between Bob and Chuck, either strong or weak. When that tie does not exist, it is known as the Forbidden Triad.

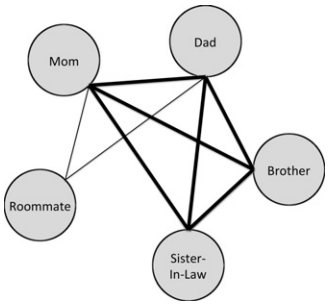


FIGURE 5.1

Sample Exercise. Note that there are strong ties connecting four of the five people listed, and two more weak ties. Only two ties are absent, between the roommate and brother, and roommate and sister-in-law. This is a very densely connected network with many strong ties.

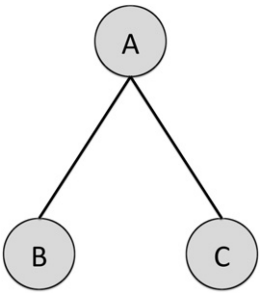


FIGURE 5.2

The Forbidden Triad.

Granovetter named this triad "forbidden" because of the unlikelihood that no connection between Bob and Chuck exists. It is an exaggeration to say this never occurs, but studies have shown that it occurs less frequently than one would expect if tie strength were randomly distributed between people in the network. One can also think of this structure as representing something actually forbidden, like a person (A) who is married (to B) and is also having an affair (with C).

From this, a second principle relating network structure and tie strength arises. Consider Figure 5.3.

In this network, P and F have a strong tie connecting them. This is the only edge that connects F's cluster of nodes to P's cluster of nodes, so it is a bridge. Recall that a bridge is an edge that is the only connection between two groups of nodes.

Nodes P and F have other strong ties as well. Indeed, in almost all social networks, nodes have more than one strong tie. In this network, F has a strong tie to O, and P has strong ties to H and N. We can form three triads with P, F, and another node where there are two strong ties: PFO, PFH, and PFN. In these cases, if there is no third connection, we are left with a forbidden triad. For example, we expect that there should be an edge between P and O since strong ties connect F

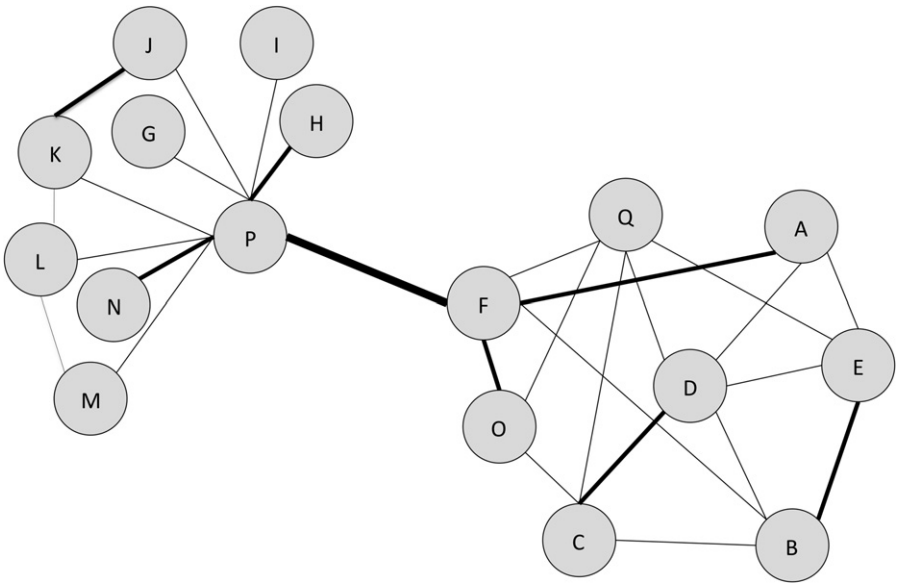


FIGURE 5.3

The edge between P and F is a bridge that connects the two clusters of nodes. It is a strong tie, and thus we would expect connections between some of the triads with two strong ties (e.g. PFO, PFH, PFN). It is very unlikely that no tie third tie would exist in any of those triads, and thus it is unlikely that a strong tie would be a bridge.

to both P and O. If such an edge were to exist, even as a weak tie, then the edge between P and F would no longer be a bridge; the new edge (e.g., an edge between P and O) would be another path connecting the two clusters. Thus, since several forbidden triads are unexpected, it is very unlikely that a strong tie will ever be a bridge; another edge is likely in one of these triads, and that will add another connection between the clusters. Granovetter described this in his work with the principle that *no strong tie is a bridge*; while strong ties *may* be bridges, it is unlikely given what we know about the distribution of edges. It is also unlikely that, over time, a strong tie would remain a bridge. Weak ties would be likely to form and connect nodes to remove the strong tie's bridge status.

Tie strength and network propagation

Network propagation is a phenomenon where things spread through a network. These may be diseases spreading through a social network, computer viruses on the Internet, or rumors and fads through a social network. Later in this book, we will delve deeper into the topic of network propagation, but for now we will discuss specifically how tie strength relates to the rate at which phenomena can spread through a network.

For something to spread from one person to another in a network, there needs to be a path between them. This can be a direct edge, either a strong or weak tie, or a series of edges between mutual acquaintances. It will pass from one person to some of his or her neighbors, and from them on to their neighbors.

Consider a disease that infects person B in the network shown in Figure 5.3. It can be passed from B to C, D, E, and F. Person B will not necessarily pass it to all of his neighbors; he may not see some of them or simply may not infect others. For this example, say B passes the disease to C and F. F can pass it to A, O, P, and Q and C can pass it to D, O, and Q. In this second phase, let's say O, P, and Q are all infected. Then they can pass it to their neighbors and so on. This describes the propagation of the disease through the network. The same reasoning could apply to a rumor, a viral video, or a piece of news.

Many factors play in to how things propagate through networks, and tie strength is one of them. As we saw in Figure 5.1, a person's strong ties tend to be connected to one another, often by more strong ties. Granovetter proposes that as the strength of a tie becomes stronger, the overlap in social circles will tend to increase. This means that if we follow all edges from our strong ties, we will re-encounter many of our own friends. That in turn implies that if we pass information to our strong ties and if they pass it to their strong ties, it will not go very far; instead, it is likely to reach people who have already received the message.

If we pass that same information to our weak ties, it has a chance to go farther in the network. Since there is usually smaller overlap in friends between a person and their weak ties, the weakly connected people have a chance to spread the information to new people whom the source did not know.

This is the effect Granovetter found in his study about finding jobs. Because a person has more weak ties than strong ties, and because the weak ties are connected to diverse social groups that the person would not otherwise communicate with, the weak ties are more often sources of new information, like job opportunities.

Weak ties are powerful in helping spread information farther through a network, but this is not to say that strong ties are unimportant. Strong ties tend to be more trustworthy, reliable, and personal. A weak tie may be able to tell a person about many job opportunities, but a strong tie will have a better idea of what jobs would be a good fit. Strong ties are also more likely to do things to help one another (i.e., reciprocal services).

Exercises

1. Define the term *tie strength*.
2. Give an example of an interaction between two people that falls into one of the areas labeled A,B, C, D, or E in the box. For example, roommates in college may spend a lot of time together, and because they are living in the same space, they may have many intimate interactions.
3. What are Granovetter's four factors that influence tie strength? List each and give an example of each
4. You have been sent to a public place—a park, the mall, or the like. Without listening in on conversations, your task is to observe pairs or groups of people and guess their tie strength.

   What observable things could you measure to determine this? For example, two people may have a conversation and you could measure how long each person talks. That could tell us if one person dominates the conversation, which may be informative. As another example, you may count the number of times people touch one another or the distance they stand from one another.

   For each of the seven factors relating to tie strength—Intensity, Intimacy, Time, Reciprocal Services, Structural, Emotional Support and Social Distance—list three things you could measure from observing people interact in public. Explain why each would be useful for understanding tie strength.
5. Find a public discussion group online. You can find many of these on the Yahoo! Groups website, but consider looking for one related to topics that interest you (favorite band, pets, sports team, video game, etc.).
   Read the posts to develop an understanding of the type of discussions happening, the most active people, and their interactions.
   a. Repeat the analysis from Exercise 3: For each of the seven factors relating to tie strength—Intensity, Intimacy, Time, Reciprocal Services, Structural, Emotional Support, and Social Distance—list three things you could measure from observing people interact in the forum. Explain why each would be useful for understanding tie strength.

b. Test your theories from part a. Make a list of five of the most active people on the discussion board. Group them into pairs (e.g., Person 1 with Person 2, Person 1 with Person 3, and so on). There will be 20 pairs. Based on your reading of the conversations, estimate the tie strength between each pair. Then, for each pair, look at their interactions and measure the things you listed in part a. Record your observations. How well do the interactions you recorded reflect the actual tie strength?

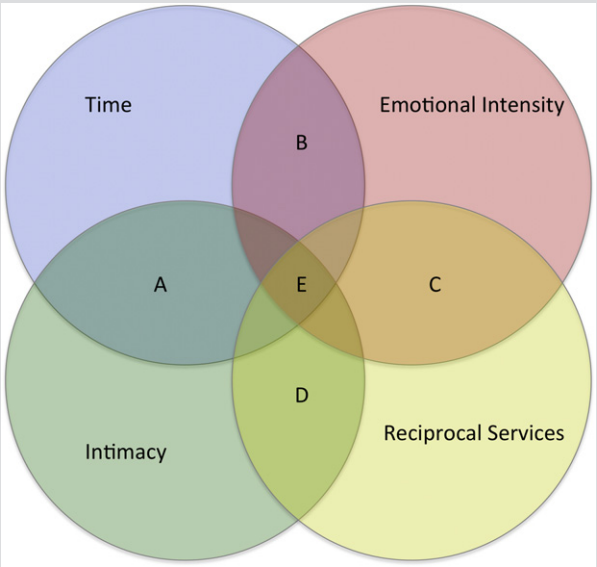
6. Give an example of the role weak ties play in job search.

7. Besides job search, list three tasks where a person might rely on weak ties to give them important information? Then, list three tasks where strong ties might be better information sources. Explain why for each.

WHEN DO FACTORS MATTER?

The figure below shows Granovetter's four original factors related to tie strength. For each factor, describe a situation where having a strong relationship with that factor will lead to a stronger tie. Then describe where a strong relationship on that factor does *not* result in a stronger tie. Do the same for each of the labeled overlaps. For example, with label A, think about a relationship where people spend a lot of time together and have a high level of intimacy.

(Note that overlaps between Time and Reciprocal Services and between Intimacy and Emotional Intensity are not pictured here, but they can also occur.)



Can you see any patterns among situations where a tie is not strengthened despite having many of these factors in place?

This page intentionally left blank

CHAPTER

Trust 6

Defining trust

Trust is a relationship with which we are all familiar, but which we rarely define or describe. Examples of trust are easy to produce, and that can be helpful in coming up with a general definition of trust. Consider these questions.

What is something you would do with a person you trust that you would not do with someone whom you do not trust? Write down a few examples for yourself. Common answers include:

- Loan money.
- Ask for a movie/restaurant/hotel recommendation.
- Tell a secret.
- Ask for a recommendation letter or for a reference.
- Ask for advice.
- Ask the person to take care of my pets/house while I am away.
- Lend my car.
- Go on a trip.

By claiming these as examples of things we do with a person we trust, it is implicit that we expect the interaction to go well. For each item you listed and each of the examples above, consider how we expect the person to react. Here are a few examples to go with some items from the list above.

- Loan money—We expect the person will pay us back
- Ask for a movie recommendation—The person's recommendation will match our taste, and the movie (or restaurant or hotel) will be good.
- Tell a secret—The person will keep a secret, not tell others, and not judge us for it.
- Ask for a recommendation or reference—The recommendation will be positive and help us get the position we are applying to.

Each of these examples shows that, when we trust someone and have one of these interactions, we expect the person will do the right thing and that it will lead to a good outcome.

This is a fundamental part of a trust relationship. The person being trusted is expected to do the "right" thing. This usually means she will act with the other person's best interests in mind and/or take actions that benefit the other person.

The person making the decision about whether or not to trust someone is considering more than just her expectations about the other person's actions. She must also decide if she is willing to take some risk by putting her trust in the other person. That may be a small risk or a large one. Receiving and acting on a poor movie recommendation may only waste a few hours of time, but making a large loan that is not repaid can have major effects. So can asking for a recommendation letter from someone who will not write a good one.

All of these ideas can be condensed down into several important factors. First, the person doing the trusting must make herself vulnerable and take some risk by trusting the other person. Second, she takes that risk because she believes the other person will act well or behave in a way that will benefit her. Vulnerability, risk, and positive expectations of the other person are the core components of the trust relationship.

Thus, as a definition of trust, we can say the following: *A person trusts another if she is willing to take a risk based on her expectation that the trusted person's actions will lead to a positive outcome.*

This definition has all the major components found in many other definitions put forth by sociologists. Deutsch (1962) and Golembiewski and McConkie (1975) use a frequently referenced definition of trust. They state that trusting behavior occurs when a person encounters a situation where she perceives an ambiguous path. The result of following the path can be good or bad, and the occurrence of the good or bad result is contingent on the action of another person. If the person chooses to go down the path, she has made a trusting choice. Sztompka (1999) presents and justifies a simple, general definition of trust similar to that of Deutsch: "Trust is a bet about the future contingent actions of others." The bet represents the vulnerability and risk, and because it is based on the future action of others, it captures the part of our definition about having positive expectations of the trusted person.

Note how these definitions differentiate some types of beliefs from trust. For example, someone may think that her favorite sports team will win the championship, but simply believing that does not mean that she *trusts* them to win. The vulnerability and risk components are not present, and so her belief alone is not enough to qualify as trust.

Nuances of trust

Trust is a dynamic and complex part of human interaction. The definition and examples discussed above are intuitive, but there are several important issues about trust that should be discussed. This section will address issues about how trust is built, how it relates to context, and how it changes over time.

Development of trust

Trust is formed between people in a wide variety of ways. In a common scenario for building trust, one person develops trust in another over time through a series of interactions that help the person build up a belief in the reliability and good intentions of the other, eventually to the point where she is willing to take a risk and act on the building trust. A series of risks that are rewarded lead to more trust. However, this does not always happen and is often not possible. Consider meeting a physician for the first time. Someone may trust that doctor with his health, but it is based on factors such as background, qualifications, references from other people, and personal compatibility rather than on a series of successful interactions over time. Someone may also have to immediately develop trust in another person, such as a victim trusting a rescuer in an emergency.

McKnight et al. (1998) document four major components of the way people consider and build trust in others:

1. Calculation-based trust: This is a rational decision about whether to trust someone, and where the costs and benefits of trusting are factored in.
2. Personal-based trust: This reflects a person's propensity to trust, developed over the course of their life.
3. Cognition-based trust: This describes the instant rapport and trust that can develop between people who share similar backgrounds, beliefs, and values. It often is based on first impressions.
4. Institution-based trust: This addresses how trust may form in the presence of guarantees and protections offered by an institution.

These factors can be applied in a wide range of contexts, including the study of trust between people and from people to organizations or communities.

Asymmetry

For two people involved in a relationship, trust is not necessarily identical in both directions. Because individuals have different experiences, psychological backgrounds, and histories, two people may trust each other at different levels. For example, parents and children clearly trust one another differently. Children must have almost absolute trust in their parents, while the parents may have almost no trust in their children, particularly when they are very young.

This strong asymmetry can occur in other relationships where the people involved are on close social levels. This can be carried out fully to "one-way trust" where circumstances force one person to trust the other, but there is no reciprocal trust (Hardin, 2002; Cook, 2001). However, most asymmetry is not as extreme as any of those circumstances. Most trust is mutual (Hardin, 2002) in that each party has some trust for the other, but often there are still differences in how much they trust one another. For example, employees typically say they trust their supervisors more than the supervisors trust the employees. This is seen in a variety of hierarchies (Yaniv and Kleinberger, 2000).

Context and time

Except for a few of these very asymmetric relationships, like that between parents and young children, trust is rarely all-encompassing. Rather, a person will tend to trust someone else about a set of things, but not about everything. For example, someone may trust her friend to recommend a movie but not to repair her car. She may trust her boss to edit a document but not to perform surgery on her. When people build trust in one another, and when they rely on it, it is usually connected to those contexts. However, trust may sometimes transfer from one context to another. A person may build trust in a co-worker that is entirely in the work context, but later trust that person to recommend a plumber, even if they have never had a discussion about plumbing or household repair.

Trust can vary from one context to another, but even within a given context, it can change over time. As discussed above, people tend to develop trust over time as their history of shared experiences builds. Trust may also decrease. If someone has one dramatic failure, trust may disappear completely. Someone may recommend a restaurant that her friend hates. That could result in a sharp decrease in the trust that the friend has in her.

Measuring trust

Measuring trust is important but difficult. People perceive trust differently, and trust is also difficult to quantify or explain. When studying how to measure trust, it has generally been broken down into two parts: a person's propensity to trust, and one person's decision about the other person. We will look at each of those factors independently.

Propensity to trust

Some people are more trusting than others. One person may trust a stranger quite readily, while someone else may be cynical and take a long time to win over. These inclinations are personal and vary independently of who the person is deciding to trust. A person's tendency to trust others is referred to as their propensity to trust.

A common way to measure a person's propensity to trust—and the trustworthiness of others—is through the Investment Game (Berg et al., 1995). In this game, a person must decide how much money to invest with someone. The basic game works as follows. The *sender* receives some initial money, say $10, to start. He must then decide how much of that to invest with the *receiver*. The *receiver* is given three times what the sender invests. For example, if the sender decides to invest $5, the receiver is given $15. The receiver then decides how much of that money to keep, and how much to give back to the sender as a return on his investment.

If people are completely self-interested, the sender would never invest money. The receiver makes the most money if he keeps everything and gives nothing back to the sender. If the sender understands this, he realizes that he will lose anything he invests, and thus he will not invest anything. However, when this game is actually played, senders do tend to invest, and receivers tend to return some of the money to the sender. Some results (Berg et al., 1995) show that on average, senders tend to get a 95% return on their investment.

From this game, we can conclude that senders who invest more money are more trusting and that receivers who return more money are more trustworthy.

Of course, some senders choose not to invest, and some receivers choose not to return any money to the senders or to return less than was invested. The actions of both players can give insight into how trusting the sender is and how trustworthy the receiver is. It reveals a person's value for his or her self-interest and commitment to ideas of fairness, moral behavior, and the interests of others.

There are also surveys that measure the propensity to trust. Evans and Revelle (2008) created a survey based on work by Goldberg (2006) that determines person's propensity to trust and trustworthiness by asking her to rate how accurately statements describe her. You can take this test yourself in the box in this chapter, and compare your scores to the average scores that the researchers found in their work.

In general, the questions gauge the test−taker's belief in the goodness of other people, interest in the plight of others, belief in strictly enforcing rules, and value of cooperation and positive social interaction. These values parallel those measured in the Investment Game, where senders and receivers demonstrate through their actions how much they believe the other person will act fairly and how much they value generosity over self-interest.

Trust in others

Based on an individual's propensity to trust, the amount of trust he has in others will vary. However, trust is not entirely dictated by the truster's personality and trusting nature. The truster's beliefs about the specific other person they are deciding to trust are very important.

Recall that the definition of trust includes the truster making himself vulnerable and taking a risk based on his expectation of the other person. A person's propensity to trust measures his overall willingness to take risks and overall expectations of people to generally behave well. Decisions about others deal with the risks a person is willing to take with a specific other person.

The survey in the box includes questions not just about a person's propensity to trust, but also trustworthiness. A trustworthy person acts respectfully and with consideration to the needs of other people, tries to be fair and act in line with established rules and expectations, and is honest and reliable. This test will help you determine your own trustworthiness and compare it to others. The questions about trustworthiness are also things to consider when determining the

trustworthiness of a specific person. It is important to consider an individual's trustworthiness in addition to how trusting a person is to people overall.

Trust is multifaceted, and a number of facets are common factors that influence how much one person trusts another. Johnson-George and Swap (1982) categorize these as follows:

1. Trust with material possessions
2. Belief about reliability
3. Trust with secrets
4. Trust regarding physical safety

Trust with material possessions captures how much someone would trust another with objects or money. This is a consideration when you are deciding to loan something—a book, a car, money, and the like. A person who is trustworthy with material possessions will take care of the object and return it on time and in good condition. More generally, trustworthy people will be fair to others and act within expectations and rules.

Belief about reliability generally describes how much people will do what they say they will do. This includes keeping appointments and showing up on time, fulfilling obligations, keeping promises, and so on.

Trust with secrets means that a person is trusted to hold a confidence. This means the confidant will not reveal secrets to other people, nor will he harshly judge the secret teller. These are people who will keep confidences and who may also offer good advice about them.

Trust regarding physical safety usually refers to how much someone trusts another to have his or her physical best interests in mind. The truster believes the other person will not harm her physically. It also means that the trusted person will have the truster's best interests in mind. For example, if the truster were to have an accident, the trusted person would do her best to help, protect the truster, and make decisions that were in line with what the truster would want. Similarly, patients trust doctors to advise correctly about their physical health and to do the right thing to protect that health when the patient is in a vulnerable position. This can also be extended to trusting a person with respect to the physical safety of others. Parents trust babysitters with the physical safety of their children, for example.

To help quantify trust in these domains, the researchers mentioned above presented a survey that measures overall trust, which includes trust about material possessions and safety, emotional trust, and reliability. The survey presents a series of statements that the test-takers rate on a scale indicating how strongly they agree or disagree. Examples include:

- Overall Trust
- If we decided to meet somewhere for lunch, I would be certain _____ would be there.
- I could expect _____ to tell me the truth.

- Emotional Trust
- I could talk freely to ____ and know he/she would want to listen.
- _____ would never intentionally misrepresent my point of view to others.
- Reliability
- If my alarm clock was broken and I asked _____ to call me at a certain time, I could count on receiving the call.
- If I were injured or hurt, I could depend on _____ to do what was best for me.

Trust in social media

Online, it is much harder to judge a stranger's trustworthiness. The information available about a person is much smaller, forging an identity is easier to do, and there is often no shared history or past interaction on which to rely. Furthermore, the number of people with whom it is possible to interact is vast; instead of being limited to one's social circle or people in the physical area, it is possible to interact with nearly anyone of the billions of people online.

Before social media and user interaction became the dominant paradigm for web use, users were mostly concerned with how much to trust websites. E-commerce in particular presented a challenge, and users generally did not trust businesses with their financial information or their personal information (McKnight et al., 2000). Online retailers worked to overcome this by creating privacy policies, by offering assurances about the safety and security of transactions (sometimes certified by third-party security websites), by building websites that looked professional and worked well, and by developing and using encrypted protocols (e.g., HTTPS).

These efforts were all designed to address the same kinds of trust issues that people have with trusting other people. Privacy policies show the website's commitment to keeping secrets. Guarantees of transaction security and safety show they are trustworthy with material possessions. Professional websites that function well are designed to prove that the company is reliable.

Once users began interacting with one another online, concerns about websites' trustworthiness combined with concerns about people's trustworthiness. One of the earliest examples of these two concerns coinciding is eBay. The online auction site was founded in 1995 and grew quickly to hosting 2 million auctions in 1997 in the early days of e-commerce. On the site, sellers list items for auction, potential buyers bid, and at the close of the auction, the buyer pays the seller who then ships the item.

This holds a lot of risk. Buyers pay before they see the item, so they risk receiving a bad item or receiving nothing from the seller. To help address this problem, eBay added a *reputation system*. Buyers and sellers rated one another after each transaction. Someone could see other users' history of feedback to help them make a decision about whether or not they were trustworthy.

Since its introduction, eBay has made many changes to its reputation system. Many other sites where users interact with one another directly also have reputation systems in place. Although they all work in slightly different ways, the core idea remains the same: Users provide feedback about one another to help others make decisions about trust.

In the 2000s, social media began to grow, and by 2010 it was the dominant way people were using the web. Static web pages maintained by web professionals were no longer the most common content online. Blogs, social networks, social bookmarking systems (like del.icio.us or Pinterest), and video websites began to produce huge amounts of content. This shift made the web a place that was no longer just human-to-website interaction, but a place for human-to-human interaction.

People participating in social media share a lot of personal information. Some of that is intended for public consumption and is not sensitive, but other information is private and should not be shared. Users must trust both the website and other users to treat their personal information with respect. Privacy policies can address some issues of trust in websites, though often social media sites are liberal about sharing and users are not aware of the implications of their policies. All the issues of trust in other people come into play online as well, and they are compounded by some of the same factors that were at play in eBay: Users do not necessarily know everyone who will have access to their information, they often have no history on which to build trust in those people, and even knowing a person's real identity is difficult.

Furthermore, reputation systems usually do not exist in these types of applications. There are many interesting and complex issues that relate to trust in social media, based on what users want to share, with whom they want to share it, and what control they have over those decisions. These factors and our understanding of them are constantly changing. This book will discuss some of the issues in greater depth in the chapter on online privacy.

Inferring trust

A problem that frequently arises online is that one person wants to know how much to trust another. Trust between two people in a social network can be considered a weight on the edge that connects them, much like was discussed with tie strength. A natural question that follows is to ask how that number can be obtained.

As mentioned earlier, the two major considerations that impact trust are an individual's propensity to trust and that individual's decisions about how trustworthy the other person is. The surveys included in both sections could be administered to people, and their results combined to generate a trust rating. However, social networks are large, and most people will not take the time to fill out a

survey about every person they know. On Facebook, for example, where users have hundreds of friends, completing a survey about each one would be a daunting task. Even if users simply rated their trust in others (e.g., on a 1−10 scale), it would be a lot of work to add those ratings for each friend and to maintain the scores as trust changes.

Furthermore, a person will often want to know how much to trust a stranger. In that case, there is no personal history on which to take a survey or to provide a rating. Offline, people may ask their friends or friends of friends for information about the stranger's trustworthiness, but online, a stranger may be very socially distant (e.g., a buyer may want to know how much to trust a new seller on eBay), and finding the people to ask about trustworthiness can be a lot of work.

Thus, a method that can estimate how much one person will trust another will be very useful. Fortunately, the problem of inferring trust has been widely studied, and a number of methods are available for it. These techniques are still cutting-edge research that is being refined, but they provide insight into ways trust can be computed online.

Network-based inference

Assume people have rated their trust in their friends. It could be everyone they know, or a subset of people they know best. Our social network will have nodes for each person and contain all the edges with trust ratings.

Consider Figure 6.1. This shows a social network with trust values on the edges, rated from 1 (low trust) to 10 (high trust). For simplicity in this example, the network is shown as undirected. However, since trust can be asymmetric, it would be appropriate to use a directed network with different values in each direction.

Consider the problem that the node "Source" wants to know how much to trust the node "Sink." *Source* and *sink* are common terms to use when talking about pairs of nodes in a graph, and they refer to the start and end node, respectively. The source does not know the sink directly, but there are many paths through the network. Some are shorter, like the path directly through node A. Others are longer, like Source - B - D - G - Sink. The trust of individuals along those paths varies, too. Source - B - E - I - Sink has several low-trust values, while Source - B - D - F - Sink does not. Only F's trust in the sink is low in that case.

How can the source use all the information on these paths to come up with a guess of how trustworthy the sink is? The social network can offer insights. A trusted friend of a trusted friend may himself be trustworthy. In Figure 6.1, the source has relatively high trust in node B, who in turn has high trust in node D. Thus, the sink could reasonably conclude that node D is also relatively trustworthy.

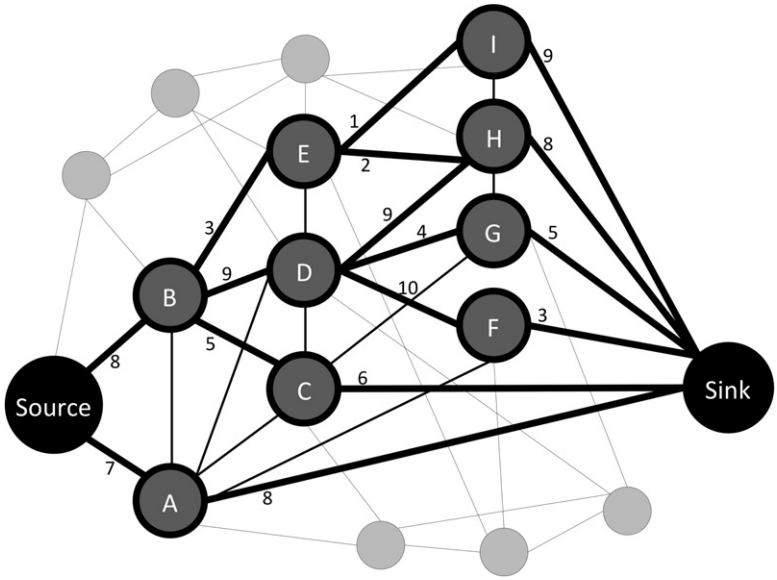


FIGURE 6.1

A social network with trust values shown as number weights on some of the edges. Trust is rated on a scale from 1 to 10 where 1 is low trust and 10 is high trust.

As paths get longer and trust values vary more, this logic becomes less and less reliable. A friend of a friend of a friend of a friend is a more tenuous connection, even if there is high trust along every step of the path. Thus, the best way to leverage the information in the network is to favor highly trusted connections and short paths over long ones.

Computer scientists have built many algorithms for using the trust values in the social network to determine how much the source will trust the sink. Some look only at the direct trust ratings of the sink and use those to come up with a single value of the sink's overall trustworthiness. These are called *global* trust algorithms because they come up with one global trust rating for the node in question. In this example, averaging the trust ratings from nodes A, C, F, G, H, and I to the sink would be a simple way to do this. Another method may consider the trustworthiness of A, C, F, G, H, and I (based on other users' trust ratings of those individuals), and use that to adjust how much their ratings of the sink are considered. Some of these methods are similar to Centrality measures discussed earlier in this book.

Other algorithms compute a personalized trust value depending on which node is the source. The idea behind this is that one node may have high trust, while another may have low trust. As context, consider politics. If someone wants to know about whether a candidate for office is trustworthy, the answer depends on

who is asking. Some people will share the candidate's views, and then trust will be high. Others will have very different views and trust will be lower. Algorithms that compute these personalized values are called *local* trust algorithms.

There are many local trust algorithms. Here is one example of how they work. The source asks its neighbors about how trustworthy the sink is. Some will know directly (like node A), and they will tell the sink their value. Others will not know, so they will ask their friends. Their friends will know or not, and the ones who do not will ask their friends. This goes on until a series of paths are found to the sink. These are bolded in the network in Figure 6.1. Once a node has information from its neighbors about the sink, it combines their values. For example, node D has ratings from nodes H, G, and F. Node D could simply average those values, or it could weight the information from each node based on how much it trusts the node. In that case, node D would give more weight to nodes H and F because they have high trust, while node G would get less weight. Once a node has calculated a trust estimate from its neighbors, it can report that value back to any other node who asked for a trust value for the sink. Those values get passed back eventually to the source.

Experiments have shown that many of these algorithms can estimate trust with high accuracy, often to within roughly 10% of the actual value people would assign to one another.

Similarity-based trust inference

As mentioned above, people often do not supply trust ratings in social networks. Thus, there are also methods for inferring trust based on other data in the network. Research has shown that people who trust one another tend to be similar (Ziegler and Golbeck, 2007). A person will trust his friend about movies if they have similar taste, or a parent will trust a babysitter to watch her child if they have similar ideas about the appropriate way to care for the child and respond in an emergency.

Other research has demonstrated that similarity is an important component of trust, but more nuanced factors are also at play (Golbeck, 2009). In particular, if there is a major disagreement between people, or if they disagree (even moderately) on an issue that is very important to one of the people, trust may be low. For example, if two people agree on every political issue except whether abortion should be legal, and that is the most important issue to both of them, their trust for one another about politics could be very low, even though they are very similar overall.

When people have provided ratings of items (e.g., they have rated movies), the similarities in their ratings can be used to estimate trust. Giving more weight to big disagreements or disagreements on items that are very important to someone (e.g., items they have given the highest or lowest possible rating to) can improve the estimate. Research has shown that computing trust based on

similarity in ratings can be about as effective as the network-based methods, getting to within around 10% of the actual trust ratings people would assign one another (Golbeck, 2009).

Researchers are actively working on solving problems such as these. Their approaches vary widely and often consider all information that is available about people—their connections in a social network, ratings they have given to items, profile information they provide in social networks, patterns of how they use social media sites, their history of interaction with one another, and so on. We are on the cutting edge of understanding how to use information from the web to understand trust relationships, just as we are with computing tie strength, as was presented in an earlier chapter (Gilbert and Karahalios, 2010). There is a lot of potential for trust, tie strength, and other relationships to be inferable from online information. Early results show it can be done, but there is still much to be done before a full picture of the issues, traits, and complexities of this problem is available.

Exercises

1. Give three examples of where the trust one person has in another will increase very quickly. Give three more examples where trust will decrease quickly.
2. Think of some ways you could vary the Investment Game to learn more about a person's propensity to trust. What if you changed the amount of money or the number of times the sender and receiver played together? What if you had the sender and receiver change roles after the first round and play again? Describe a variation you come up with and what you think it might reveal about trustworthiness.
3. Four aspects of trust are listed above: trust with material possessions, reliability, secrets, and physical safety. The survey in the box has statements that touch on some of these. Come up withfive5 new statements that could be added to a similar survey. Which of the four aspects of trust do they relate to?
4. Get a small group of your friends together (five to eight people). Have them each take the trust test in the sidebar and compute their scores. Then have them play the investment game with one another. You do not need to use real money.
   a. Do the people who have higher trust scores on the sidebar test tend to invest more money?
   b. Do people with higher trustworthiness scores on the sidebar test tend to have more money invested with them?
   c. Do people with higher trustworthiness scores on the sidebar test tend to return more money in the investment game?
5. Look at the graph in Figure 6.1 with trust values on the edges. A simple global trust algorithm is to use the average of the trust ratings assigned to a node as its trustworthiness. Compute the trustworthiness for nodes A through I and for the sink.

6. Consider trust in your email.
   a. Look at the last 50 messages you have received. Make a list of all the people who have emailed you in that time.
   b. Rate the trust you have in each of those people. You may rate trust with respect to a particular context (e.g., professional trust) or give a more general trust rating.
   c. Independently, rate the importance of each email as low, medium, or high.
   d. How well does the trust you have in the email senders relate to the email importance? Do low-importance emails tend to come from people you trust less?
   e. Using Gephi or NodeXL, build a network of your email contacts. Add in the trust ratings you created as weights on the edges. Create a visualization that uses those edge weights.
   f. Analyze the network you created in (e), with attention to the trust relationships. Do more trusted people tend to be connected? Is there any pattern in the trust relationships?
7. Imagine you are in charge of creating a new way to sort the messages a person sees in his social media feed (Facebook news feed, Twitter feed, etc.). You have access to the entire social network. Also assume that users have the opportunity to rate how much they trust their friends (but not all will do so).
   a. How will you use trust to sort the news feed?
   b. How will you deal with cases where people have not rated the trustworthiness of their friends? How can you estimate trust in those people?
8. Look at the survey in the callout. Recall that there are four categories of trust they are measuring: trust with material possessions, belief about reliability, trust with secrets, and trust regarding physical safety. Think of three additional statements for each of those categories that you could add to the survey in the box. For each, explain why you think it is a good statement to test trust in that category.

HOW TRUSTING AND TRUSTWORTHY ARE YOU?

This test was designed by Evans and Revelle (2008) to test how trusting and trustworthy a person is. Try it yourself here and see how you compare to others.

Directions: Below, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex and roughly your same age.

Please use the following scale as you are answering the questions:

(1) Very Inaccurate
(2) Moderately Inaccurate

(3) Slightly Inaccurate
(4) Slightly Accurate
(5) Moderately Accurate
(6) Very Accurate

	Statement	Your Response
1	Listen to my conscience	
2	Anticipate the needs of others	
3	Respect others	
4	Can get along with most people	
5	Have always been completely fair to others	
6	Stick to the rules	
7	Believe that laws should be strictly enforced	
8	Have a good word for everyone	
9	Value cooperation over competition	
10	Return extra change when a cashier makes a mistake	
11	Would never cheat on my taxes	
12	Follow through with my plans	
13	Believe that people are basically moral	
14	Finish what I start	
15	Retreat from others	
16	Am filled with doubts about things	
17	Feel short-changed in life	
18	Avoid contacts with others	
19	Believe that most people would lie to get ahead	
20	Find it hard to forgive others	
21	Believe that people seldom tell you the whole story	

How Do You Compare?

Scoring: The questions here cover your propensity to trust (how trusting you are) and your trustworthiness. To get your score for each attribute, add up your ratings from above as instructed below.
Comparing: The average trust score found by researchers was 40.8. The average trustworthiness score was 44.3. Compare yourself to the rest of the population by drawing a vertical line on the charts to indicate your position.

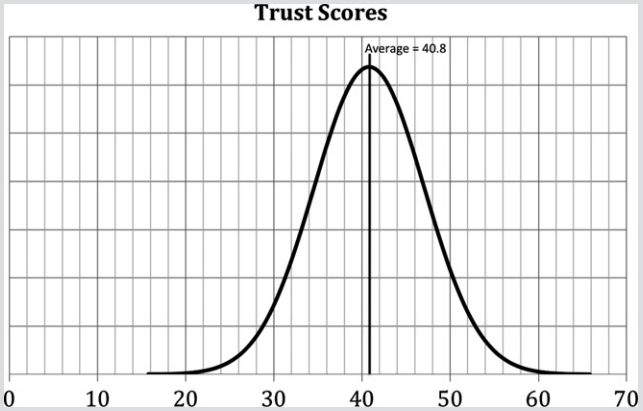
Propensity to trust

Add your values from questions 4, 8, 9, 13.

For questions, 15–21, reverse your scores by taking 7 minus the rating you gave (a score of 6 becomes a 1, a score of 1 becomes a 6, and so on).

Add those values to your total. That is your trust score. Enter it in the box below.

Trust Score: _____

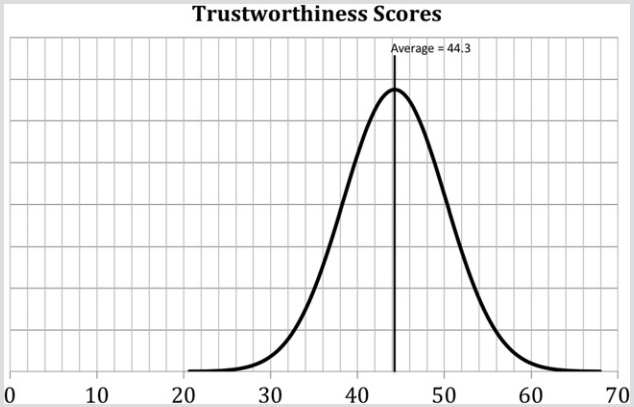


Trustworthiness

Add your values from questions 1, 2, 3, 5, 6, 7, 10, 11, 12, and 14.

That is your trustworthiness score. Enter it in the box below.

Trustworthiness Score: _____



This page intentionally left blank

CHAPTER

7 Understanding Structure Through User Attributes and Behavior

Structural analysis of social networks provides insight about how people relate to one another and where they fit within the larger pattern of connections. *Structure* only reveals some of what is happening in a network. After understanding *patterns* of relationships, it's necessary to delve deeper and analyze attributes of the people and their interactions in the network.

To really understand what's happening in a network, however, it is important to look beyond the structure—at users' attributes, behavior, the content they are sharing, and their interactions. Analyzing this content can lead to many insights into the meaning of network structure.

Attributes, *behavior*, and *content* are terms used throughout this chapter. *Attributes* are characteristics of a node. For a person, attributes would include age, gender, or location. They could also be preferences or beliefs, like a person's religion or political preferences. *Behavior* refers to actions of nodes. For example, behavior of a person in a social network could include how frequently she posts, who and how many people she follows, or how often she clicks on shared links. *Content* is a broader term that refers to the nonstructural information about a node. It includes any information about the nodes or edges beyond the structural features discussed earlier in the text. It may also include information related to nodes that is a combination of attributes and behavior, like the types of comments they post online or the topics they discuss. If a node represents a video or picture instead of a person, the content would include the things depicted in the video.

As an example of analysis using content, consider the network in Figure 7.1. There are three clear clusters of nodes. Structural analysis could identify those clusters and provide statistics about the network properties (like density and connectivity), and about the importance of individual nodes with measures (like centrality).

But what is the meaning of these three clusters? Is there some feature that nodes in the same cluster share in common? Do the clusters' attributes hold some information that could provide insight into the overall network? This requires analyzing more information about the nodes and their attributes.

The network in Figure 7.1 is built from the photo-sharing website Flickr. On Flickr, users upload photos and label them with descriptive keywords called *tags*.

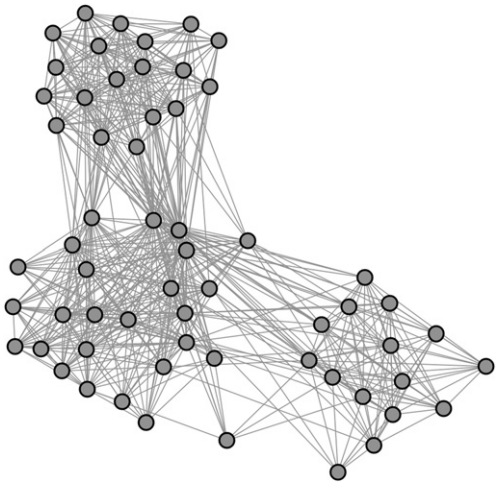


FIGURE 7.1

A network with three clear clusters. What do they mean?

This network is created using those tags. The nodes represent tags, and an edge between tags indicates that they were used to describe the same image. For example, if an image is tagged with the words "desk" and "keyboard," the network would show a line connecting those two words. A tag network like the one shown is the 1.5 egocentric network of a specific tag. Any tags used on the same images as the central tag are included as nodes. Then, edges are added to link any of those tags that were used together on at least one image.

Looking at Figure 7.1, the network structure indicates that the central tag is related to three groups of other tags, but it doesn't explain why. What do the tags have in common? What do the clusters mean?

These tags were all used on images that were also tagged with the word "mouse." "Mouse" is the tag used to generate this 1.5 ego network. This may lead to theories about the clusters, but to truly understand the phenomenon, one must look at the tags themselves. Figure 7.2 shows the same network with nodes represented by tags instead of circles: One cluster is about a mouse (the animal), one cluster is about a computer mouse, and one is about the character Mickey Mouse.

Analyzing the attributes of the nodes—in this case, the label—reveals valuable insight about the clusters and the network. As another example, consider Figure 7.3. In this graph, there are two obvious clusters: one on the left, and one on the right.

This network is built from Twitter. It is the 1.5 egocentric network of a user. The nodes are people who follow or are followed by the user. Edges indicate a following relationship between them. From the structure, it is clear that this user

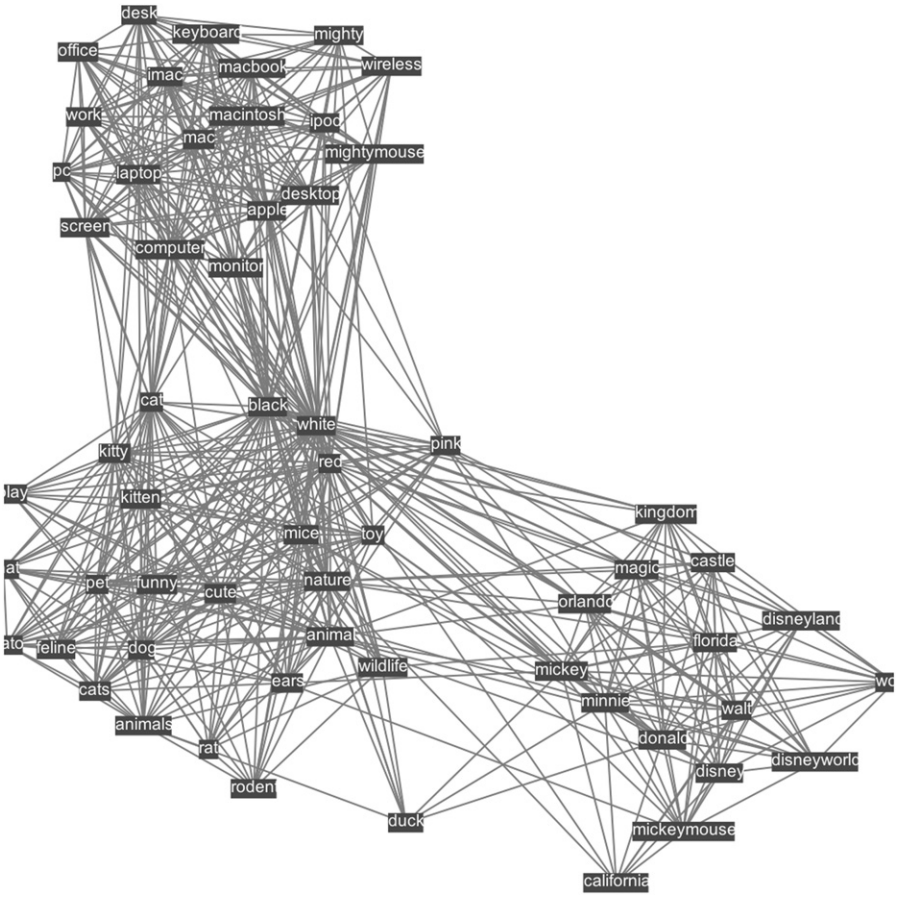


FIGURE 7.2

The same network as Figure 7.1, this time shown with the tags that the nodes represent. The network is built of tags used with the tag "mouse," and the three clusters have clear themes representing a computer mouse, an animal mouse, and Mickey Mouse.

communicates with two groups of people that are largely distinct from each other, but the graph does not reveal who is in each group or why they are separate.

The color of the nodes represents the primary language used on Twitter. The black nodes are people who primarily post in Spanish, and the white nodes post in English. There are also gray nodes found throughout the graph, and these people use multiple languages. As with the example above, understanding the attributes of the nodes provides much more insight than the structure alone.

User attributes require different levels of analysis to discern. In the example from Figure 7.1, merely seeing the tags revealed enough information to

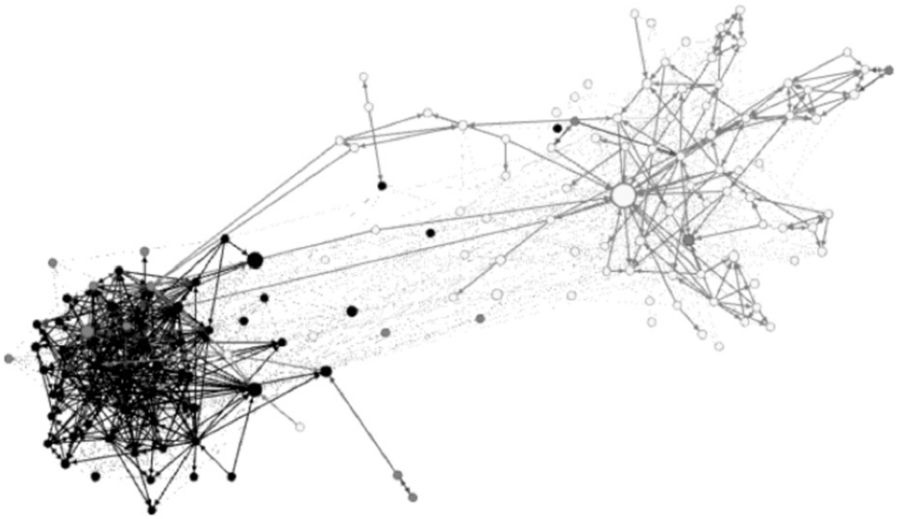


FIGURE 7.3

The 1.5 egocentric network of a Twitter user. There are two obvious clusters. Black-colored nodes post primarily in Spanish, and white nodes post only in English.

understand the patterns in the network. In Figure 7.3, each user's primary language attribute had to be determined by reading some of their tweets. Language can be considered a *behavior* of the users, as well as an attribute. Behavioral features can be helpful in understanding users in many ways.

Figure 7.4 shows another 1.5 egocentric network gathered from Twitter. Once again, notice the visible clusters. But in this network, demographic information about the users does not explain what is happening. All users speak English, and secondary languages are unrelated to the structure. People in the different clusters cover the same age ranges, gender, and education level.

What else can we look at to understand the meaning of the clusters? The users' behavior holds the key. On Twitter, behavior consists primarily of posting content. Thus, if we look at what people post, it tells us a lot about how they are using the service. Reading sample tweets from people chosen out of each cluster shows that people in the tight cluster on the left tend to tweet about the Washington Capitals, the professional hockey team in Washington, D.C. People in the larger, looser cluster to the right tend to tweet about academic and social media issues. There's also a small group of connected people in the upper left; these users generally tweet personal messages, often related to fashion or pop culture. These groups represent communities related to the central user's different life contexts as hockey fan, social media researcher, and friend.

These examples demonstrate how attribute and behavioral data can be useful for gaining deeper understanding of a network's structural features. The rest of

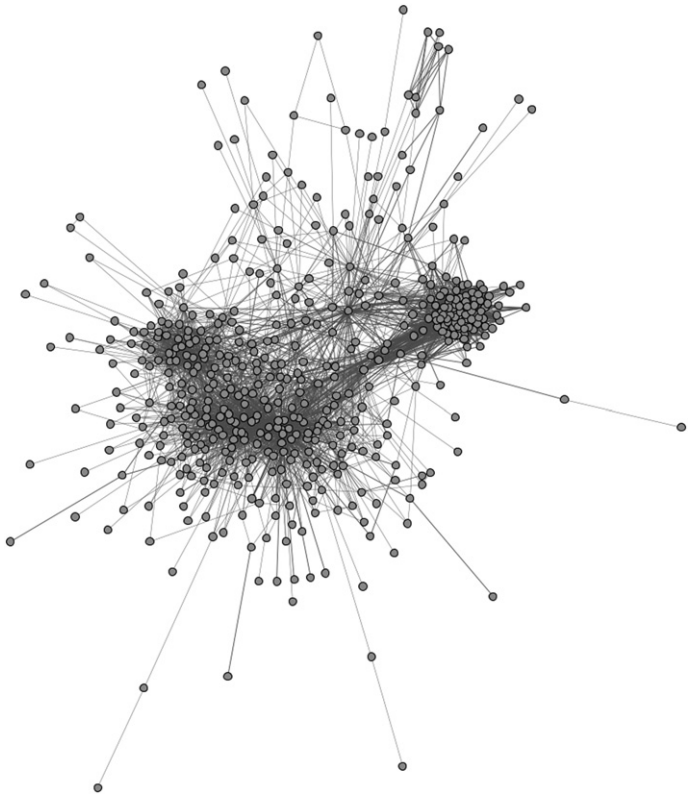


FIGURE 7.4

The 1.5 egocentric network of a Twitter user.

this chapter will focus on guidelines for conducting this type of analysis, and illustrate its application in a case study that identifies user roles in online communities.

Analyzing attributes and behavior

The examples presented above dealt with clustering patterns in networks, but this is just one structural feature that content analysis can explain. Analyzing user attributes and behavior may provide insights into almost any network feature.

For example, a network may have many singleton nodes (or small dyads and triads), disconnected from the main component. These disconnected nodes may all have attributes in common. Or they may lack an attribute that unites the majority of users in the connected component.

Actually, this commonly occurs in social networks. Often, there's a subset of users who register for the site, explore it once, and never return. These users often make no connections during their one visit and thus remain disconnected from the group. This behavioral attribute explains the structural feature of many singletons.

A network may also have patterns of structure. There may be some users who have a very high out-degree, but low in-degree; others with a very low degree; and still others who appear to have many strong connections to other users. Is there some common attribute or behavior among people who share similar structures? Sometimes there is. The case study in this chapter illustrates this in depth.

Analyzing content

This section outlines a process for analyzing networks by using content. The first and most important step is to understand the context in which the network arose. If it's an email network, be familiar with the emails. Read many to develop a sense of the people, the topics discussed, and the purpose of the messages. If the network is built from a social media website, read user profiles and their posts. If it comes from a discussion board, read many posts. While it may sometimes be possible to simply guess at what attributes will help explain structural features in the network, it's most often an understanding of the people in the network and their behavior that leads to ideas for combining structure and content.

After becoming familiar with these basics, the next step is to visualize the network. Look for patterns. These may include clusters, isolated nodes, or recurring structures. For example, there may be some hubs, some low-degree nodes, and some nodes with very strong connections. Perhaps users with similar structural attributes share other things in common. Often, there will be some characteristics that stand out in a visualization, and these are an ideal place to begin analysis.

Once you have structures of interest, the next thing to do is probe the content to try to explain the structures. For example, if there are several clusters, randomly select some nodes from each cluster. What do nodes within a cluster have in common? Look at their attributes (e.g., gender, age, language spoken, etc.) and their behavior. Are people in one cluster behaving similarly to one another, but different from people in other clusters? Do shared structural attributes indicate common personal or behavioral attributes among people? These questions will probably follow directly from a good understanding of the network's context, users, and their general behavior.

After examining different attributes and behaviors in terms of their relationship to the network structure, the next step is to validate the relationship. Different visual properties of the nodes such as color, opacity, and shape can be used to represent attribute and behavioral data such as gender, age, and language. As shown in Figure 7.2, labels can also be effectively used to represent content, particularly for smaller networks. Does a clear pattern emerge? Do the majority of nodes in a cluster share the attributes you have identified? Do the majority of nodes with a similar structure share the attributes that you hypothesize explain it?

You will almost certainly not find perfect agreement between structure and these traits, but, as with the examples above, the pattern should be clear.

Example analysis

To try the techniques described above, consider the network in Figure 7.5.

The network is built from YouTube. Each node represents a video that was tagged with the keyword "cubs." Edges link videos that share at least one additional keyword in common.

Two large clusters are clearly defined in this network: to the left and right. To understand what those clusters represent, we select sample videos from each cluster. Figure 7.6 shows the five random videos selected from each group. Those from the cluster on the left are highlighted in white, and the ones from the cluster on the right are highlighted in black.

The next step is to learn about each video, discover its attributes, and determine what it may share in common with other videos in its cluster. YouTube provides a lot of information about each video. Table 7.1 shows a subset of keywords (or all keywords, if the list was short) for each video. Reading the keywords shows that all videos highlighted in white from the cluster on the left are about the Chicago Cubs, a Major League Baseball team. Videos in black from the cluster on the right are about animal cubs—lions, tigers, bears, and cheetahs.

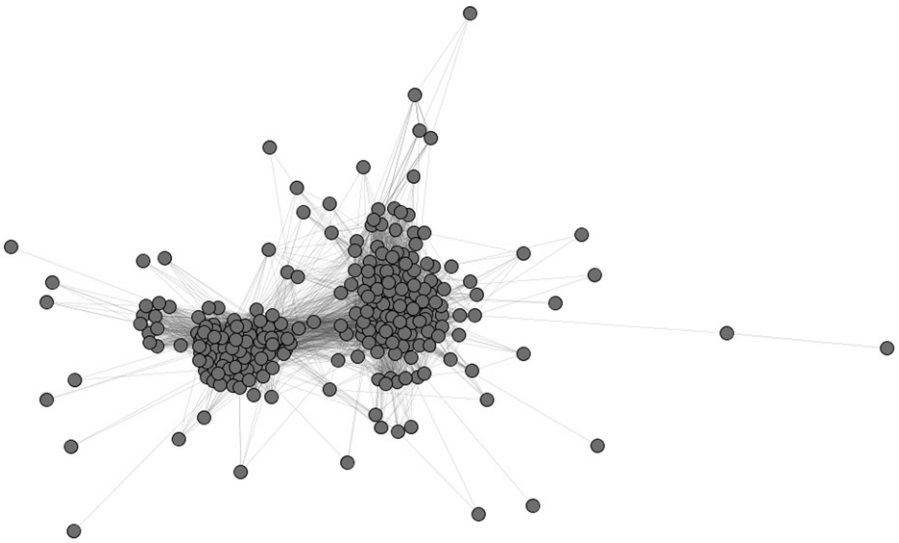


FIGURE 7.5

A sample network with two clusters. The nodes represent YouTube videos. Edges link videos that have been tagged with the same keyword. All videos were tagged with the keyword "cubs."

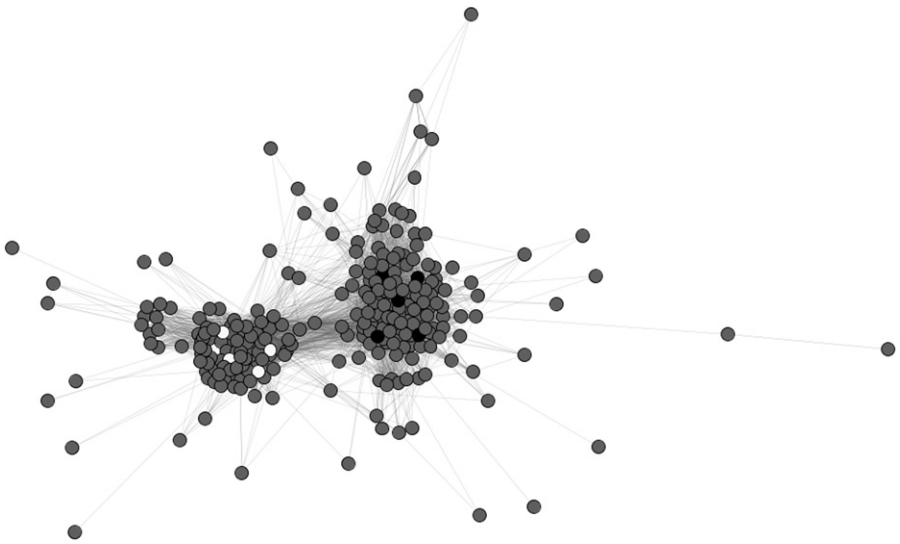


FIGURE 7.6

Selected nodes from each cluster highlighted in white and black in the graph.

Identifying that there are two different topics within the main clusters is a major insight into this network. Having done that, additional structural analysis can continue to provide useful information. For example, the density of the Chicago Cubs cluster on the left is 0.52, while the density of the animal cluster on the right is 0.34. The baseball videos have 1.5 times as many links. Since links represent shared tags, this could mean that there is a tag or small set of tags that is extremely common on those videos, whereas such a tag does not exist for the animals. A closer analysis of the edge data, which lists the tag shared between videos, shows that either "Chicago" or "Chicago Cubs" appeared on 1,975 of the 2,984 edges in that cluster, accounting for roughly ⅔ of the edges. While there were common tags in the animal cluster, too (e.g., "zoo" and "cute"), none were as popular or as important as the "Chicago" tag was to the baseball cluster.

This example illustrates how structural analysis can inspire questions that can only be answered by looking at the attributes of the nodes and edges. In turn, this can lead to more structural questions and more content-based questions. The result is a much deeper understanding of what is happening in the network than one could achieve by using any single type of analysis alone.

Case study: Identifying user roles

The baseball example above is a relatively simple example that is useful for explaining the basic process of analyzing attributes, behavior, and content in relation to the structure of a network. In this section, we present the results of some

Table 7.1 Keywords for each of the Sampled Videos

White 1	White 2	White 3	White 4	White 5	Black 1	Black 2	Black 3	Black 4	Black 5
Cubs	mlb	MLB	Chicago Cubs	Chicago	dog	National Geographic	Tiger	tiger	cheetah
CubFans	2k12	12	Chicago	Cubs	dogs	polar	Rescue	tigress	cheetahs
baseball	baseball	The Show	Cubs	Spring	puppies	bear	Lions	cubs	african
Chicago	major	MLB 2k12	Wrigley Field	Training	pup	cubs	Leopards	machli	wild
Please	league	Diamond Dynasty	Opening Day	Baseball	cute	mother	Cubs	fight	cute
Stop	ronnie	Baseball	2011	Tony	adorable	mom	Kittens	nick	animals
Believing	woo	triple play	number one fan	Campana	snuggle	parent	Tiger cubs	ranthamore	baby
	wilckers	world series	sports fans	Brett	bear cub	learn	Wild animal orphanage	croc	BBC
	wrigley	home run derby	baseball	Jackson	Medvjedić	teach	Big Cat Rescue	crocodile	cubs
	cubbies	PS MOVE	major leagues	Sports	Bär	cute	Texas	mugger	
	north	Jose Bautista Chicago		Hohokam	orsacchiotto	fluffy	Tigers	india	
	side	Cubs		Park	brown bear cub	sweet	Rescued	rajastan	
	billy	win		Cactus	bears	predator	Scary	valmik	
	goat	sports		League	teddy	arctic	Roar	thapar	
	curse	playstation			medo srečko	predation	Rawr	bbc	
	illinois	ps3			cubs	hunt	Attack	wildlife	
	ps3	ps vita			medvedji mladič		Aggressive		
	playstiation	video game			slovenia		Sanctuary		
	cubs	so real it's it's unreal			slovenija		Global		

fundamental research in this area where the combination of structure and content analysis leads to important discoveries about how people interact in online communities and forums.

Researchers studying Usenet, a pre-web technology for online discussions, noticed that when they visualized the networks, there were differences in the structure of users' egocentric networks. They decided to investigate if these structural differences related to the roles that users played in the online communities.

After identifying the structures, the researchers read each person's posts.

The results revealed many specific behavioral roles corresponding to unique structural features. The relationship between these roles and network structures have been reconfirmed many times in web-based discussion groups and on mailing lists (Welser, Gleave, and Smith, 2007).

Figure 7.7 shows the social network of a forum for discussing cascading style sheets, a technology for structuring web pages. The nodes in the network represent people who have participated in the discussion. The edges are directed and indicate when one person has replied to another. Interactions were collected over a three-month period.

As a whole, the network looks fairly typical. Figure 7.8 breaks the network down, showing the 1.5 egocentric networks for many users. Here, the differences between user interaction patterns are easy to see. Some people interact extensively, while others have only a few edges. Some egocentric networks are very dense and others are not. Do these differences represent clear differences in the

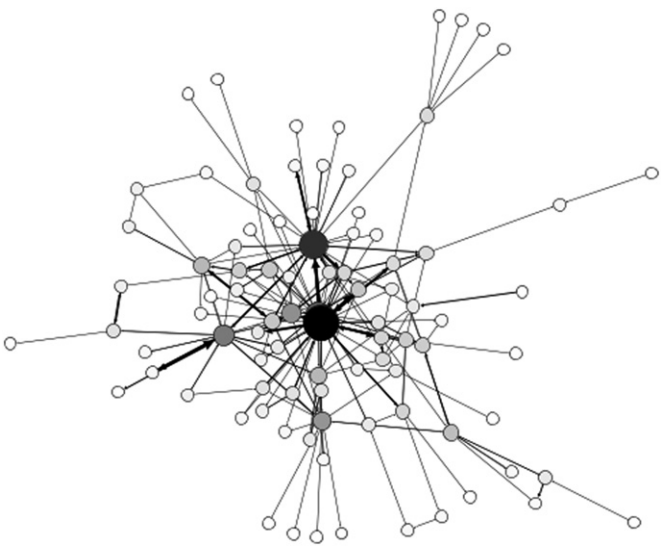


FIGURE 7.7

The network of three months of discussion on the CSS-Discuss mailing list. Node size reflects the node's out-degree in this directed network.

social roles people play in the forum, or do they simply show a range of behavior that does not correlate with any specific role? Answering that question requires examining the content of users' posts.

The number of each type of egocentric network is not representative in Figure 7.8. There are many dyads and triads, as shown in the bottom row, and far fewer nodes with higher degree or with the denser egocentric networks as shown in the top row.

Do these different egocentric network structures relate to consistent behavioral differences? As a first example, consider the egocentric network in the lower right of Figure 7.8. The central node has only one connection; a node with a high degree has replied to post from the central user. This is extremely common in this network; 36 nodes have only one neighbor, and in almost all cases that neighbor has a high degree and had replied to the central node. Another 17 nodes have two neighbors with this same pattern. This accounts for nearly 60% of the nodes in the network.

The next step is to examine the content that these users are posting, as well as content posted by the people who reply to them. In nearly every case, these nodes with a degree of 1 or 2 have posted a question, and the high-degree node has

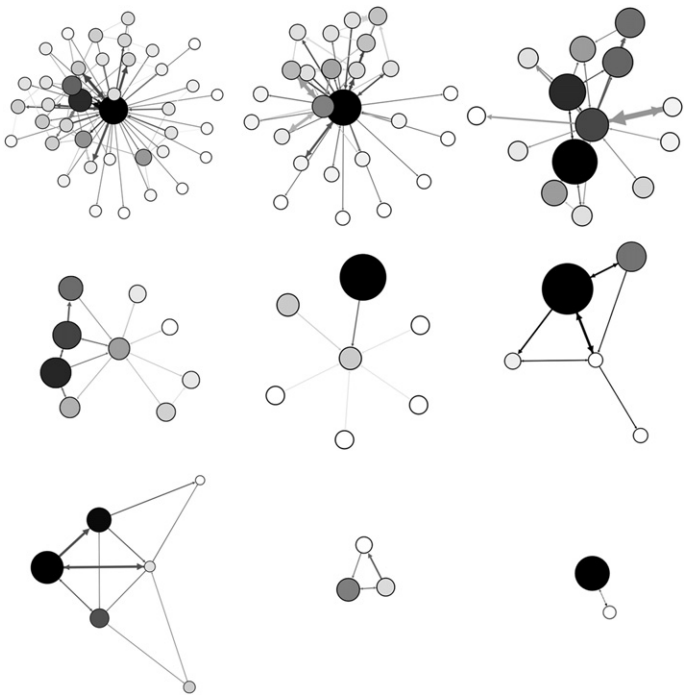


FIGURE 7.8

Sample 1.5 egocentric networks of users from the network in Figure 7.7. Both size and color indicate degree. The egocentric node is always in the center of the graph, but it may not be the largest or darkest.

answered them. These have been called "Question People" in the research on this topic. They are found in many types of communities, and have very similar patterns. They will ask a question, get a reply, and then basically stop participating in the community. Structurally, they have a low in-degree and out-degree, and their neighbors (the question answerers) tend to have a high degree.

In contrast, consider the first two networks in Figure 7.8. The central nodes here have a high degree but are connected mostly to people with a low degree. The networks are not very dense—there are some connections between neighbors, but mostly the central node has responded to people and little additional discussion is present. Reading the posts in these threads reveals these people to be the complement to the Question People. They are "Answer People" who tend to answer questions posted by others. Structurally, they have a high out-degree, their egocentric networks have a low clustering coefficient (since those they answer to don't typically reply to each other), and their neighbors have a low degree.

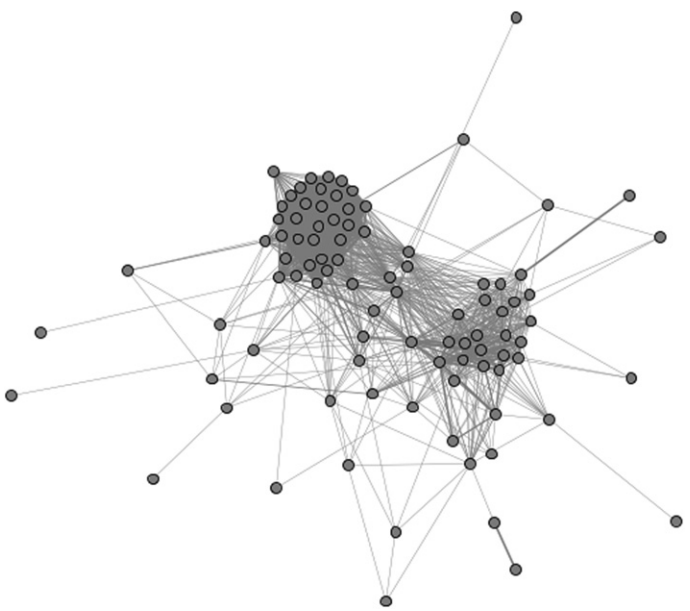
Finally, consider the rightmost graph in the top row of Figure 7.8. The central node has a relatively high degree compared with the Question People. However, unlike Answer People, the neighbors have a relatively high degree. The clustering coefficient of the egocentric network is also much higher than that of the Answer People. Reading the posts reveals that people with this structural pattern tend to be "Discussion People." They start, and sometimes participate in, discussions. Their neighbors tend to have higher degrees, and the clustering coefficient of their networks is relatively high.

Researchers have discovered other patterns representing Trolls, Spammers, and Flame Warriors (Turner et al., 2005). The relationship between egocentric network structure and its corresponding role has been identified in many different types of communities in studies conducted over many years. This kind of research serves as a good example of how identifying structural features, in combination with analyzing the behavior and attributes of users, can lead to new and interesting insights about the network being studied.

Exercises

1. Pick 10 of your friends and family members.
   a. For each person, list the following attributes:
      i. Age
      ii. Hometown
      iii. Education level (years of education)
      iv. Occupation
   b. Create three more personal attributes and list those for each of the people you chose.
2. Examine your personal email. List major categories of topics you discuss with people. These may include topics like work, classes, family, or social events. Come up with a list of 5−10 major categories of topic.

3. Look at your email and choose the 10 people with whom you think you correspond the most.
   a. For each person, read the last 10 messages you have exchanged with him or her. Which of the major categories from question 2 apply to your conversations? Indicate each person and the categories of discussion you have with them.
   b. Create a network of these people, adding edges between anyone who has been on the same email message together.
   c. In the network, is there any pattern in the network structure that corresponds to the high-level categories?
4. Create a visualization of your entire personal email network. This can be done automatically with many network visualization tools, like Gephi or NodeXL, or you can do it by hand. Let the nodes represent people and let the edges indicate if they have been on an email message together.
   a. Look at the network. Are there any clusters or features that stand out?
   b. Select an attribute that you think may define separate groups of people you communicate with. This may be the context in which you know them (from high school, college, work, activities, etc.), their age, or other factors. Color or mark the nodes to indicate the attribute that describes each. Do these attributes help explain patterns in the data?
5. The network below is a result of searching YouTube for videos tagged with the word "bunny." (You can make a similar visualization for yourself using NodeXL and the YouTube spigot). Links indicate videos that share at least one other tag in common.

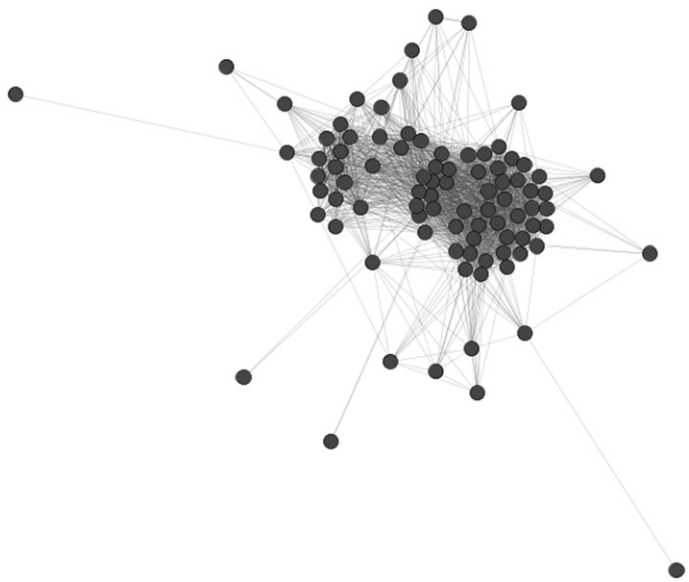


   a. Identify all the structural attributes of interest. What are the clusters? Which parts of the network are dense or sparse? Which nodes seem remarkable (important, outliers, etc.)?
   b. We have selected three videos each from the cluster in the upper left (Rq3MGlzC5l8, 10kL5oOiHRk, FFuitd30vH4) and the lower right (wSFB2ytWJLQ, hgDHWLyztCI, 1SqBdS0XkV4). Search for those codes on YouTube to see the actual videos and the tags associated with each. Does the content of the videos reveal the meaning of the different clusters? Does it explain differences in the structure?

6. Use the NodeXL Flickr spigot to import a Related Tags network. Choose a tag that you think will produce two or more distinct clusters as was shown in some of the examples above.
   a. Before creating the graph, identify your tag and explain why you think there will be multiple clusters.
   b. Create the graph. Does it look like you expected? If there are clusters, do they represent what you expected? If there are not distinct clusters, what type of content is connected in unexpected ways?

7. Repeat exercise 6 but using a YouTube video network with a keyword that you think will produce multiple clusters.

8. The graph below is a YouTube video network for the keyword "solo." Download this data from the book website and generate a graph visualization. Explain the clusters. What feature defines the large group on the right and what defines the smaller group on the right?



9. Open the Senate Voting Records dataset in NodeXL or Gephi. Use the percentage of votes in common as an edge weight.
   a. Color each node by its political party. Visualize the graph. Do you see any patterns?
   b. Filter the edges by weight. Try a variety of weights until you feel like you have an interesting graph.
      i. Show the visualized graph.
      ii. What weight did you choose as a filter?
      iii. What is shown in the filtered graph?
      iv. What are the visible clusters, and what do they represent?
10. Analyze the Twitter account @LuvMyDogs5.
   a. Build the 1.5 egocentric network among people followed by @LuvMyDogs5.
   b. Visualize the graph of the network.
   c. Analyze the nodes in the graph. Look at their attributes. Develop a categorization scheme for the nodes.
   d. Color-code the nodes according to your scheme developed in part c. Create a new visualization.
   e. Using your understanding of the node attributes and your visualization, write a two-paragraph description of what is interesting about the structure and meaning of this network.

This page intentionally left blank

CHAPTER

Building Networks

8

There are many challenges to building adjacency lists or models to represent networks. Defining nodes and edges, making choices about what to include, and deciding how to sample from a large network are all important issues. This chapter will introduce methods for creating networks for analysis.

Modeling networks

All networks are made up of nodes and edges. Which nodes to include and what constitutes an edge are often tricky questions. Here, we introduce how to scope these definitions and make choices when building networks.

Defining nodes

Node types

In social networks, the nodes in the network usually represent people. Other types of networks have a single type of entity represented by a node. A network representing the financial industry would link banks and other companies that do business with one another. A network of an ecological system may link species based on trophic relationships (what eats what). In all these cases, there is one type of node: people, companies, species.

There are also heterogeneous networks where there are different types of nodes in the same network. For example, a network of corporate boards may have board members connected to the companies on whose boards they sit. This resulting network has nodes for people and nodes for companies. The same is true for terrorist networks which may have terrorists connected to their organizations. A wiki network may connect people to the pages they have authored. Again, there are two types of nodes—in this case, people and pages.

In mathematics, networks with two types of nodes are called *bipartite graphs* if there are edges that connect nodes of different types, but no edges connecting nodes of the same type. For example, if corporate board members are linked only to companies and there are no edges directly between the people, it would be a bipartite graph. These types of networks have a distinctive look—nodes of each type can be separated, and the edges only go between the groups; there are no edges within them. Figure 8.1 shows an example. Note that there are no direct connections between people or between organizations in this graph.

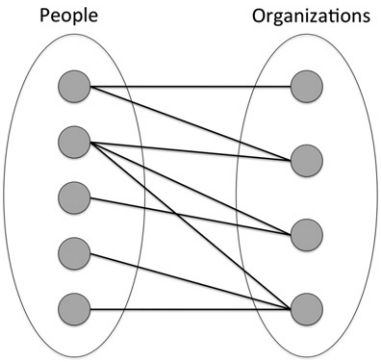


FIGURE 8.1

A bipartite graph has two types of nodes (people and organizations in this example), and edges always connect a node from one group to a node from the other group.

Other terms used to refer to these types of network are *bimodal networks* or *multimodal networks*. These terms describe networks with two or more types of nodes, respectively. When the networks connect people to organizations, it may be referred to as an *affiliation network*.

Bipartite graphs are just one example of heterogeneous graphs. There can be graphs with more than two types of nodes. For example, a terrorist network may have terrorists, terrorist organizations, countries, and government officials. Furthermore, heterogeneous graphs may have connections between all types of nodes, not just between types. A network may have edges that connect people to other people, people to organizations, and organizations to organizations.

Similarly, networks may have multiple types of edges. For example, friendships and family relationships could be included. Networks with multiple edge types are called *multiplex networks.*

This book focuses mainly on homogeneous networks with one type of node and one type of edge. Analyzing heterogeneous networks is more difficult. Many of the network analysis methods covered so far have less meaning in heterogeneous networks. For example, computing or understanding Centrality is not as clear when there are two types of nodes. Clusters may naturally end up forming around one type of node (e.g., clusters of people connected to a single organization), which then makes it difficult to see any other strong relationships in the graph.

Node selection

Once the type or types of nodes to include in a network is selected, there is still work to be done. Not every node should necessarily be included. Defining what qualifies a node to be included in a network is an important step in network creation.

For example, consider building a social network of people who work for Company X. Clearly, the full-time employees should be included. Which of the following groups would you also include?

- Part-time employees
- Contractors who are hired to come in and work temporarily for Company X for a few weeks
- People from other companies who come and work at Company , but are paid by their home company (e.g., a security company monitoring the grounds at Company X)
- People who do business with Company X but who do not work there and are not employed there (e.g., vendors who sell products to Company X)
- People who work for companies who do business with Company X but who are not involved directly in any transactions

There is no right answer as to which nodes should be included and which ones should not. It depends on the questions of interest and points of analysis to be conducted. It is important that the criteria for including nodes are clearly established before building the network, so that no one is left out or incorrectly included.

Defining edges

Edges represent relationships in networks, but relationships vary in strength and type, and they often change over time. Which ones should be included?

In some cases, the decision is straightforward. For example, relationships are clearly defined on many social networking sites: People are either friends or not. In these cases, a network model would have an edge between people who are connected on the site. Similarly, sometimes an edge will reflect a specific type of interaction. In a sexual contact network used to study the spread of sexually transmitted diseases, two people will have an edge between them if they had intercourse.

In many cases, however, the definition of an edge is not as clear. When analyzing people's interactions or behaviors, a range of relationship types and strengths emerge.

Consider an extension to the Company X example. Say you have selected the nodes in the Company X network to be full- and part-time employees only. Now, you must decide when to connect two nodes with an edge. Given two people with the following types of interactions, would you add an edge between them or not?

- Two people work in the same department and work closely on many projects. They spend several hours every day working together.
- One person works for another (a superior/subordinate relationship).
- Two people are in the same department. They participate in department-wide discussions on a mailing list and see one another at monthly departmental meetings, but do not work together on any projects.

- Two people are part of a large committee of people selected from across the company. They attend meetings together at which there are group discussions about the committee's business, but the two people have no overlap in other projects or job duties.
- Two people work in different units but have lunch together once or twice a week, where they talk mostly about personal matters.
- Two people are on the same official email list that broadcasts announcements to all employees.
- Two people met and chatted over a drink at the company picnic a couple years ago.

There is no single correct decision to be made about which of these relationships should be represented by edges and which should not. However, the question of which relationships should be included is complex. When deciding what circumstances constitute an edge, consider what the goals of the network analysis are, what relationships are relevant to those goals, and what thresholds of interaction qualify for an edge.

One way to help make these decisions is to be liberal about including edges in a first pass through the network, and then filter some out on a second pass. To do this, gather information about the edges, including relationship type or a weighting. This edge information may be present in the data. For example, an edge may be weighted given by the number of times people interact or by their tie strength. Edges may also have a label indicating the type of relationship (e.g., family, coworker, teammate, etc.).

Once the network is created, there are several techniques for analyzing or simplifying it. If there are multiple types of edges (a multiplex network), it may be inappropriate to calculate statistics, like centrality, since the edges represent different things. A multiplex network may be used without calculating statistics; visualizing the network may still yield interesting insights. The network may be separated into multiple networks with one type of edge being used in each. The edge types may also be converted into weights, thus reducing the network to have a single type of edge with different weights.

The next section provides several real-world examples with challenges regarding node and edge selection, and illustrates how to apply these suggestions to build a network.

Examples

Communication is a common way that networks form; when people communicate, there are connections between them, and aggregating these connections forms a network. If two people communicate, there should be an edge between them, and the strength of the edge can reflect the frequency or intensity of the communication.

In an online discussion board, the nodes are easy to identify: Anyone who posts a message becomes a node. Edges should be determined by communication, but

how should that be done? Say Alice posts a message and Bob responds. Then there should be an edge between them. But what if Dave enters the discussion and replies to Bob? There should be an edge between Bob and Dave, but should there be one between Dave and Alice since she posted the original message that started the conversation? If dozens of people reply to Alice's message, should everyone be connected to everyone else? That is a legitimate possibility, as is limiting the connections only to direct replies or to replies up a chain of messages (i.e., Alice, Bob, and Dave would all be connected to one another, but if Frank replied directly to Alice, he would not be connected to Bob or Dave).

As another example, consider people playing an online game. Nodes will represent players, but the time a player spends in the game may be used as a filter. By placing a lower threshold on the amount of time a player has spent in the game, nodes representing relatively inactive players can be dropped. For example, someone who joins and then quickly leaves may be excluded from the network. Edges are, yet again, the more difficult problem. Interaction between players could indicate that an edge should connect them. If two players talk to one another, is that enough to earn an edge? Do they need to engage more substantially, like fighting a battle together? And positive interactions should be treated differently than negative interactions, but how? Again, there are no correct and incorrect answers to this question. It depends on what kinds of questions the analyst has about the network and what the network would look like for each choice. If, for example, connecting all players who spoke to one another would make an extremely dense network, it is unlikely that there would be any interesting network features to analyze. Thus, this is likely a poor choice for defining an edge. However, a pattern of frequent communication may eliminate the density problem and allow for this type of interaction to lead to an edge between players. When weights are available to describe the edges, a lower threshold on the weight is often an ideal way to filter a network.

These examples are designed to illustrate the complexity and difficulty of choices in modeling networks. To probe this issue more deeply, we will consider a real dataset of email communications, make choices about nodes and edges, and look at the results.

Case study: The Enron email network

A famous collection of email that can be used to build a network and analyze an organization is the Enron email corpus. Enron, an energy company, filed for bankruptcy in 2001 after which it was revealed that the company engaged in extensive accounting fraud. During an investigation, the U.S. Federal Energy Regulatory Commission collected all available emails from Enron's employees. These were later made public through a Freedom of Information Act request. The collection of messages—roughly 500,000 unique emails to and from about 150 Enron executives and employees—is widely used to study email communication, including the structure of the network.

For the purpose of this case study, the guiding analytical goals will be to understand corporate communication within Enron.

The nodes in this network are people who have communicated by email. But which people should be included? There are people within Enron, companies with which Enron does business (some totally external and some that are subsidiaries of Enron), personal messages to employees from their families or receipts from e-commerce retailers, and so on. Any or all of these addresses could be included in a network. For the purposes of this exercise, only email from people with enron.com addresses will be included, so that the network will allow someone to analyze the communication network *within* the company.

Next is the choice of edges. The simple case—where one person directly emails another —adds a directed edge from the sender to the receiver. Not all emails are this simple. Sometimes there are multiple direct recipients of an email, and people are also included in the cc or bcc line. Should there be an edge from the sender to those recipients as well? And should there also be edges to connect all the recipients of an email to one another? If so, what would be the direction on that edge?

Any of these choices would be appropriate. For simplicity, we will choose to include an edge from the sender to any direct recipients, but not to anyone on the cc or bcc line. We will also leave out any edges between recipients.

Even with these choices, the resulting network is very large. This makes it difficult to visualize and analyze. In order to identify the most meaningful relationships, we filter the network based on the number of emails sent. Any lower bound can be used to filter the network. In this case, we choose to include only an edge if the sender has sent at least 100 emails to the recipient. That limits us to seeing only high-frequency communicators in the organization. Similarly, a lower bound could be placed on the percentage of emails the sender sends to the recipient. If the recipient accounts for less than the threshold percentage of messages, the edge could be ignored.

In Figure 8.2, we have also filtered the network so that only people with at least two frequent correspondents are shown. This cuts out people who have frequently communicated with only one other person. The giant component is visualized in Figure 8.2, and the large, dark = colored nodes represent the people with the highest degree—in this case, the highest number of frequent correspondence partners.

With a visualization like this, it is now possible to further analyze the network. Looking at only people within the company who frequently communicate, patterns emerge. This includes the very high-degree nodes shown in black, clusters like the one around the high-degree node in the center, long chains of frequent email partners as appear on the edges of the network, and “fans” where one person is connected to many single nodes (of which there are several to the lower left of the graph), indicating accounts that send a high volume of email to many people.

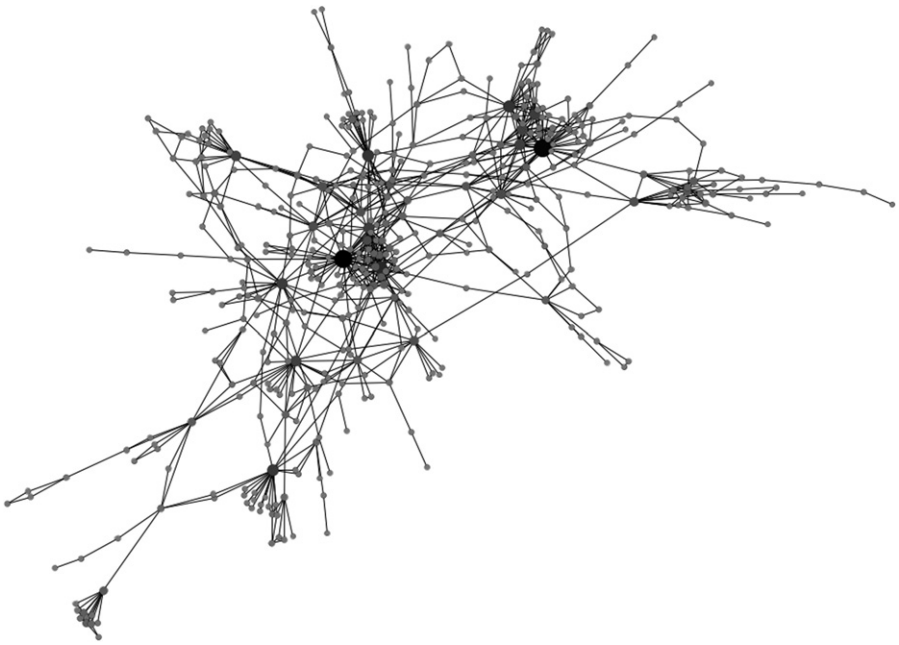


FIGURE 8.2

The giant component of the network of frequent email partners in the Enron email network. Size and color indicate high-degree nodes.

To get to the point where these interesting features emerged, there were many choices about criteria for including nodes, defining edges, and filtering the graph. This is a very important part of network analysis, since different choices could result in a very different figure and set of possible questions.

Sampling methods

After making the decisions about what nodes and edges are used in a network, the challenges do not disappear. One of the biggest issues in working with social media networks is that they are often too large to analyze in their entirety. Millions of nodes and edges are difficult to understand, impossible to visualize in a way that has any meaning, and only very general ideas about a network emerge from such a populous dataset. Sampling—selecting a subset of the nodes and edges—is an effective and common way of obtaining a reasonably sized dataset from a large one.

There are many sophisticated ways to sample a network. In this section, we present several methods that are used frequently and that are at the core of more advanced techniques. A more thorough discussion of these and other sampling

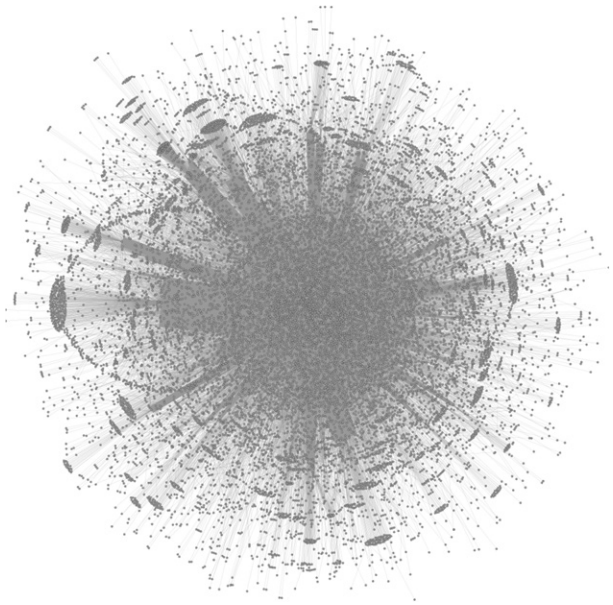


FIGURE 8.3

The Enron email network with edges connecting any pair of nodes that have exchanged at least 10 emails. Note that while some features are visible on the edges of the graph, the core of the network is far too dense to make any analysis of its structure.

methods, along with an analysis of the benefits and drawbacks of each, is available in Leskovec and Faloutsos (2006).

This section will use the Enron email network as a large graph that should be sampled. Instead of limiting the graph to people who have emailed each other 100 times as in the section above, this graph will be even larger, including edges for any nodes that have exchanged at least 10 messages. It will also include people from outside Enron. The purpose of this graph is not to better understand communication in Enron, as it was in the case study above, but rather to provide a large graph as an example and to see the impact of the sampling. The full, unsampled graph is shown in Figure 8.3.

Random sampling

Randomly sampling a network means randomly selecting a percentage of the graph to be included in a sample. The benefit of random sampling is that it reduces the network to a smaller size in an even way, so a picture of the overall patterns of relationships and clusters can be seen. However, since edges and nodes are removed from every point in the network to make the sample, random

sampling is less effective when an analyst wants to see a complete picture of the types of connections made at any point in the network.

The two most basic ways to create a random sample of a network is to randomly choose a set of edges or to randomly choose nodes.

With random edge selection, a fixed percentage of the network's edges are randomly chosen. The edge and the nodes it connects are added to the sampled graph. With this approach, the likelihood that a node is included in the sampled graph is relative to its degree. Nodes with high degree (i.e., a node with many edges) are more likely to appear than nodes with low degree, because any random edge is more likely to be connected to the higher-degree node. This introduces a bias toward higher degree nodes in the sample, which may be desirable for some analyses (e.g., identifying the most highly connected individuals).

Figure 8.4 shows the results of random edge sampling 50%, 25%, 10%, and 1% of the edges in the Enron email graph. Notice that the graph becomes increasingly sparse, but that some of the features are preserved in each sample.

Random node sampling is also an option. In this case, a subset of the nodes is randomly chosen, and then any edges that exist between the selected nodes are added into the network.

Figure 8.5 shows the result of selecting 50% of the nodes and 10% of the nodes from the same Enron email network. Notice that these networks have many fewer nodes than the networks that come from randomly selecting the corresponding number of edges.

Node sampling is not biased toward nodes with high degree, and it tends to preserve some graph statistics like the clustering coefficient and degree distribution.

Snowball sampling

Snowball sampling is a technique commonly used in sociology where participants are interviewed. In that context, interviewees are asked to refer the researchers to other people, who are then interviewed and asked for more references. The process continues until a researcher has interviewed enough people. This technique has been adapted and used for sampling large networks. To create a sample, an initial *seed* node is selected. From there, all of its neighbors are selected, and then all of their neighbors, repeating until a specified network size is achieved. The sample network grows like a snowball.

While snowball sampling is a relatively easy way to sample a network, the sample is considered biased since the nodes are all in the neighborhood of the seed node. Also, because the sampling stops when a certain number of edges or nodes are included, the network is usually full of nodes that have many neighbors who, in turn, have no connections because their neighbors were not collected. Figure 8.6 shows four snowball samples collected by starting from different randomly selected nodes in the Enron email network. Notice that there are many nodes around the edges of the visualization that have only one connection. These

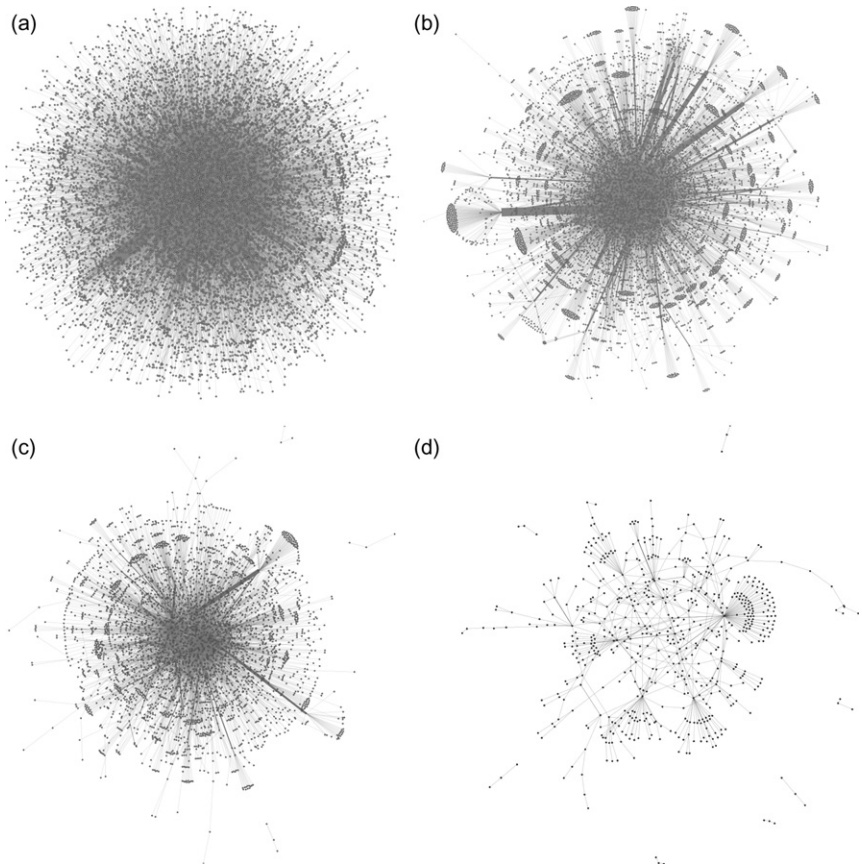


FIGURE 8.4

Results of random edge sampling on the Enron email network. The graph in (a) includes 50% of the edges, (b) includes 25% of the edges, (c) includes 10% of the edges, and (d) includes only 1% of the edges.

could be removed by using a 3.5 degree egocentric network for a snowball sample, or the nodes on the edge could be kept to indicate the size of the network that lies another step out.

Because of the bias and structural differences of a sampled network created using snowball sampling, structural statistics are not useful on these graphs. The benefits of snowball sampling come from the complete set of connections found in the core of the network. The random sampling methods above throw away information from every part of the network. If an analyst is trying to study the specific patterns of connection and clustering in a network, a random sample will be missing some of the information they want to see. A snowball sample, on the other hand, will show only a small part of the network, but it will show it

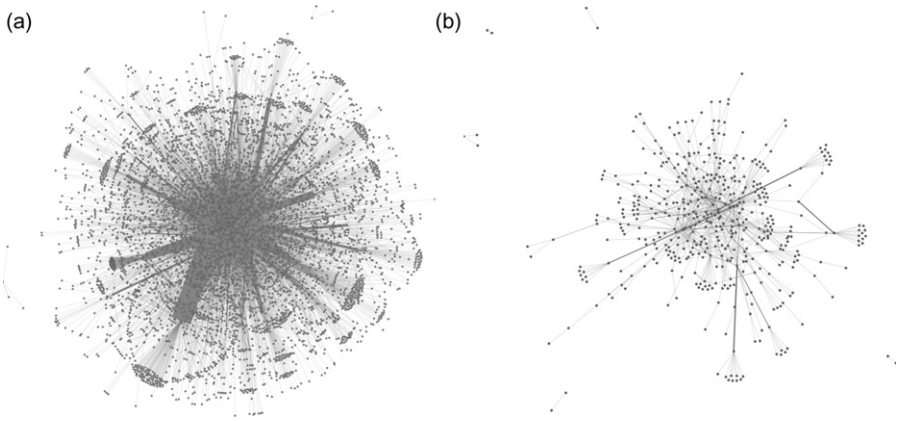


FIGURE 8.5

Networks sampled from the same Enron email network shown in Figure 8.4 (a). In this example, graph (a) includes 50% of the nodes and graph (b) includes 10%.

completely, allowing the analyst to look at the local patterns of relationships and draw conclusions. Also, at times an analyst may want to focus in on a specific subset of the network related to their research questions.

Egocentric network analysis

An alternative to sampling networks is to look at individuals in the network and analyze their egocentric networks. This method is useful both to understand features in a large network when the entire network is of interest and to select individuals who have specifically interesting traits.

When the entire network is of interest, an analyst may choose to look at randomly selected egocentric networks or the networks of individuals selected based on certain characteristics. For example, nodes with the highest centrality may be chosen and analyzed because they wield particular influence in the network.

Sometimes an analyst is interested only in individuals with particular traits. For example, in a large social networking website like Facebook, someone may want to study how high school-aged students are interacting. In that case, the structure of the overall network is mostly irrelevant. Looking at the egocentric network of many people in that age group, however, is likely to reveal a lot of relevant information.

As an example, consider the work by Eleta (2012) on multilingual use of Twitter. Her research focuses on the communication patterns of people who post messages to Twitter in more than one language.

To study this phenomenon, the entire Twitter network would have had far too much information, and most of it would have been irrelevant to the research

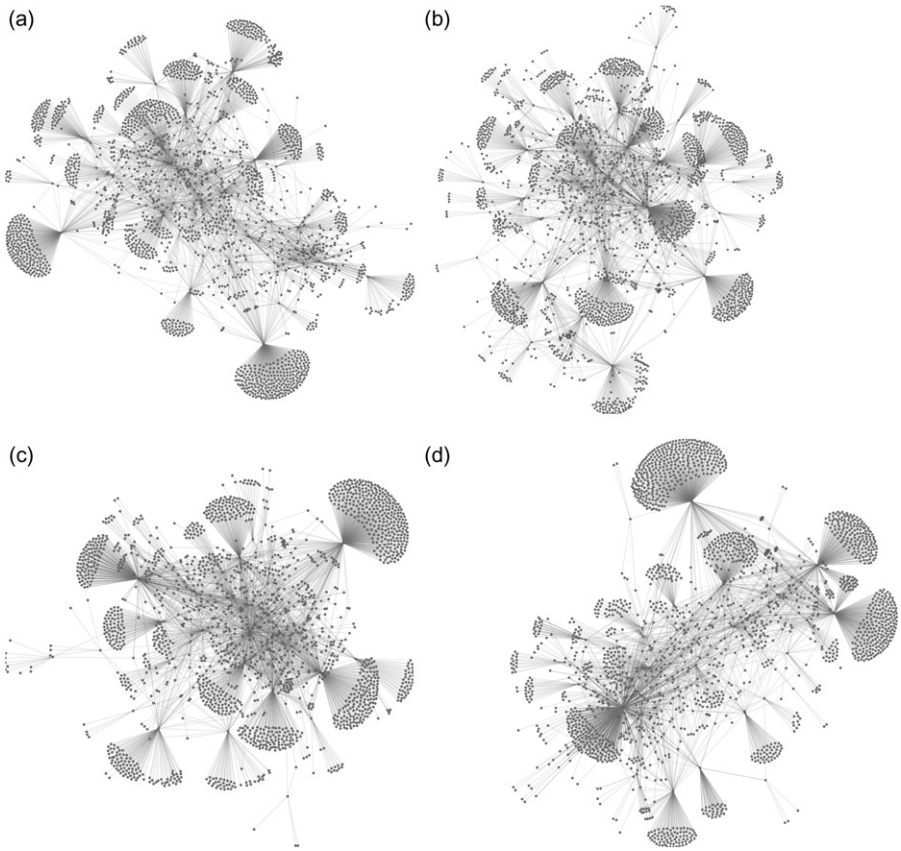


FIGURE 8.6

Four networks generated by snowball sampling, each starting from a different randomly selected node in the network. Note that all networks have large "fans" around the edges, where the neighbors of a node have been included, but those neighbors have no other connections in the network.

question. Using egocentric networks of the multilingual users provided much clearer insights.

Figures 8.7 and 8.8 show sample graphs from this research. Each graph represents the 1.5 egocentric network of a multilingual user. The original users are not included in their egocentric networks to better display the connections between their friends and followers. Color coding (black, white, and gray) in the graphs indicates the language used by each friend or follower.

Figure 8.7 is the network of a person who posts in both Greek and English. In this network of his friends and followers, those who post only in Greek are indicated in black, those who post only in English are in white, and anyone using multiple languages or a third language are in a medium gray. Two

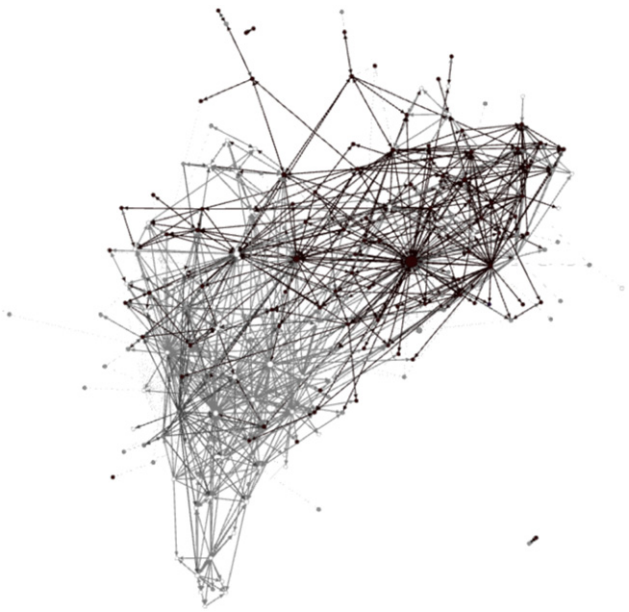


FIGURE 8.7

The 1.5 egocentric network of a Twitter user who posts in both English and Greek. Greek-speaking nodes are black, English-speaking nodes are white, and nodes using other languages or multiple languages are in gray.

patterns immediately emerge from this picture. First, the Greek speakers are clustered together and the English speakers are clustered together. However, there are many connections between the Greek and English speakers, indicating that the network contains many multilingual people even if they are only posting in one language. These users may post in a single language because they want to target a specific audience, but they consume information posted in both languages.

Figure 8.8 shows the network of a person posting in both English and Spanish. Her friends and followers who post only in Spanish are shown in black, the ones who post only in English are in white, and people using third languages or multiple languages are in gray. As in Figure 8.7, people posting in the same language tend to be clustered together. However, there are many differences between the two graphs. The cluster of Spanish speakers is much denser. There are also many fewer links between the groups. This suggests that the person at the center of this network may be reaching two separate audiences who do not communicate with one another, and who may only speak a single language. This core person would serve as an information bridge between the two groups, unlike in Figure 8.8 where there are many connections between the Greek and English speakers.

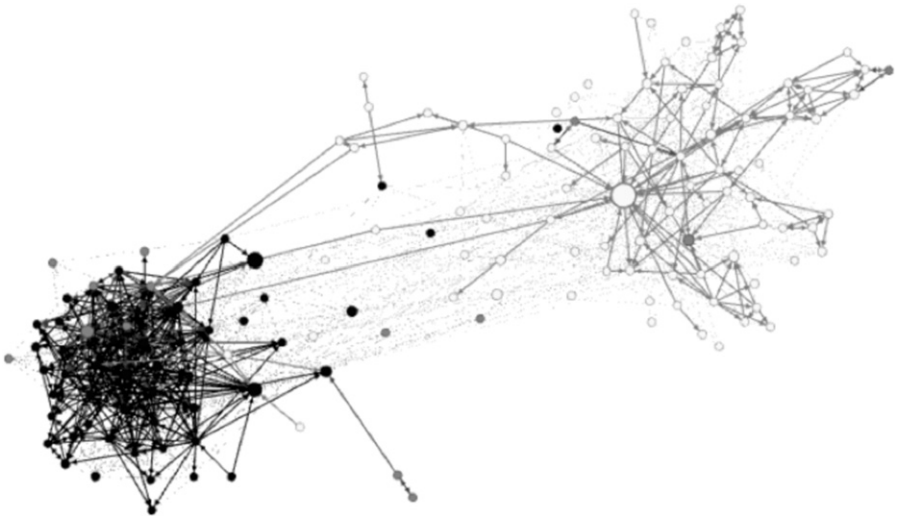


FIGURE 8.8

A 1.5 egocentric network of a person who posts in both English and Spanish. People who post only in Spanish are shown in black, those posting only in English are in white, and people using multiple languages or a third language are in gray.

This type of egocentric network analysis can reveal different types of patterns of interaction, which may in turn allow an analyst to make general conclusions about the roles, behavior, or types of nodes in the network.

Exercises

1. Create a TV Show Network.
   - **a.** Choose your favorite television show and choose a full episode. There will be many people who appear in the episode, and most will be minor characters (even without names), though the attention will focus on the main characters.
     - **i.** Make a list of all the people who appear and note roughly how long they are in the show.
     - **ii.** Use this information to set criteria for which people would qualify as nodes in a network of characters from the show.
   - **b.** What should count as an edge in your TV show network? Should it be people who interact in some way? People who appear in the same scene? People who are together for a minimum amount of time? People who have preexisting relationships (family members, friends, etc.)?
     - **i.** Define what constitutes an edge and then re-watch the show.
     - **ii.** Make an adjacency list using the characters you selected as nodes in part (a).

   iii. If appropriate, note the number of interactions, time of interaction, number of scenes, or other important information that will describe each edge.

c. Analyze the network you built in (b).
   i. Which nodes are most important and how did you measure that?
   ii. Which nodes have the highest centrality?
   iii. How does your analysis correspond with your understanding of the show?
   iv. Do your choices about nodes and edges provide an accurate picture of the social network in this episode?

2. Facebook has an obvious social network structure. Nodes are users' accounts and edges are Facebook friendships. Your goal in this exercise is to think of other networks that exist on Facebook. You may not use Facebook friendships as edges.
   a. Create at least three different definitions of networks that use people as nodes. What are the nodes and what are the edges
   b. Create at least three bimodal or multimodal networks (networks with two types of nodes). What are the nodes and what are the edges in each network?
3. Imagine you have created an awesome YouTube video that you think will go viral, and you want to track how that video is spreading through Twitter. Further imagine that Twitter has granted you access to any data you want. What kind of network would you create to study this phenomenon?
   a. What are the nodes in your network? What are the edges?
   b. How are you going to sample the network? Twitter has hundreds of millions of users, which means the entire network is too large to analyze. Which nodes and edges will be included in your dataset?
4. As you did in exercise 5 of the Tie Strength chapter, find a public online discussion board. Read the posts to develop an understanding of the type of discussions happening, the most active people, and their interactions.
   a. Imagine creating a network where the nodes represent people who participate on the discussion board. List two ways that you can define edges in this network.
   b. Imagine creating a bimodal network where nodes represent people and discussions. List two ways you can define edges in this network.
   c. Choose one of the network definitions you created in part a or b. Using the discussion board, create an adjacency list for network that meets the criteria you established. It should have at least 20 nodes.
   d. Visualize the network you created in part c.
   e. List at least three interesting network features for the network you created in part (c).
5. Find a report from your local news service that describes a complex event. It should mention at least five people, even if they are mentioned only briefly or not by name. Create a network of all people mentioned in the story. Use one node for each person. Add edges for any relationship or interaction you have information about and label each edge.

BUILD A NETWORK: EXCERPT FROM *PRIDE AND PREJUDICE*

Below is the text of Chapter 1 of Jane Austen's novel, *Pride and Prejudice*. There are many characters introduced, named and unnamed. Build a social network of the characters, labeling the edges with their relationships. This is trickier than it seems at first—some characters know about one another but have not met. Some characters are minor and others are significant. The decision of which nodes and edges to include is at your discretion, but think about your choices.

Chapter 1

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do you not want to know who has taken it?" cried his wife impatiently.

"*You* want to tell me, and I have no objection to hearing it."

This was invitation enough.

"Why, my dear, you must know, Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came down on Monday in a chaise and four to see the place, and was so much delighted with it, that he agreed with Mr. Morris immediately; that he is to take possession before Michaelmas, and some of his servants are to be in the house by the end of next week."

"What is his name?"

"Bingley."

"Is he married or single?"

"Oh! Single, my dear, to be sure! A single man of large fortune; four or five thousand a year. What a fine thing for our girls!"

"How so? How can it affect them?"

"My dear Mr. Bennet," replied his wife, "how can you be so tiresome! You must know that I am thinking of his marrying one of them."

"Is that his design in settling here?"

"Design! Nonsense, how can you talk so! But it is very likely that he *may* fall in love with one of them, and therefore you must visit him as soon as he comes."

"I see no occasion for that. You and the girls may go, or you may send them by themselves, which perhaps will be still better, for as you are as handsome as any of them, Mr. Bingley may like you the best of the party."

"My dear, you flatter me. I certainly *have* had my share of beauty, but I do not pretend to be anything extraordinary now. When a woman has five grown-up daughters, she ought to give over thinking of her own beauty."

"In such cases, a woman has not often much beauty to think of."

"But, my dear, you must indeed go and see Mr. Bingley when he comes into the neighbourhood."

"It is more than I engage for, I assure you."

"But consider your daughters. Only think what an establishment it would be for one of them. Sir William and Lady Lucas are determined to go, merely on that account, for in general, you know, they visit no newcomers. Indeed you must go, for it will be impossible for *us* to visit him if you do not."

"You are over-scrupulous, surely. I dare say Mr. Bingley will be very glad to see you; and I will send a few lines by you to assure him of my hearty consent to his marrying whichever he chooses of the girls; though I must throw in a good word for my little Lizzy."

"I desire you will do no such thing. Lizzy is not a bit better than the others; and I am sure she is not half so handsome as Jane, nor half so good-humoured as Lydia. But you are always giving *her* the preference."

"They have none of them much to recommend them," replied he; "they are all silly and ignorant like other girls; but Lizzy has something more of quickness than her sisters."

"Mr. Bennet, how *can* you abuse your own children in such a way? You take delight in vexing me. You have no compassion for my poor nerves."

"You mistake me, my dear. I have a high respect for your nerves. They are my old friends. I have heard you mention them with consideration these last twenty years at least."

"Ah, you do not know what I suffer."

"But I hope you will get over it, and live to see many young men of four thousand a year come into the neighbourhood."

"It will be no use to us, if twenty such should come, since you will not visit them."

"Depend upon it, my dear, that when there are twenty, I will visit them all."

Mr. Bennet was so odd a mixture of quick parts, sarcastic humour, reserve, and caprice, that the experience of three-and-twenty years had been insufficient to make his wife understand his character. *Her* mind was less difficult to develop. She was a woman of mean understanding, little information, and uncertain temper. When she was discontented, she fancied herself nervous. The business of her life was to get her daughters married; its solace was visiting and news.

This page intentionally left blank

CHAPTER

Entity Resolution and Link Prediction

9

Once a network is constructed, there is often missing and duplicated information. There may be multiple nodes representing the same person, or there may be missing edges. For example, consider Figure 9.1.

All pairs of nodes but one are connected in this network. While it is possible that nodes A and E do not know one another, it is extremely unlikely. *Link prediction* is a method of analysis that detects where missing links should be present in the network.

Similarly, consider Figure 9.2. In this network the nodes on the left and right—John Smith and J. Smith—have very similar names, share all the same connections, and have no connection between one another. It could be that John and J. Smith are actually the same person. *Entity resolution* is a technique for merging nodes that represent the same person, as we might do here.

Methods for doing link prediction and entity resolution can range from simple to very complex. This chapter will introduce the fundamentals of each technique and illustrate their application through several case studies.

Link prediction

Formally, the goal of link prediction is to analyze a network at a set time and predict the edges that may appear in the network in the future. However, it can be used in many applications. This can include "cleaning" a dataset. Data often has errors in it, including missing links, and link prediction could identify places where an analyst might want to check to confirm that there is no edge between a pair of nodes. It can also be used to identify people.

For example, consider Figure 9.3. This network shows how often Alice, Bob, Chuck, and Frank attend meetings together. The weight on the edges indicates the number of meetings each pair of users has attended together. If we know that Alice and Bob attended a meeting with a third person, but we do not know who the third person is, we can consult the graph to make a guess about what links are likely missing in the graph of the new meeting. While it is possible that the third person is Frank or someone not pictured, the graph in Figure 9.3 suggests that it is most likely Chuck.

Many of the network features discussed so far in this book come into play when considering link prediction. Figure 9.1 above illustrates a simple case where

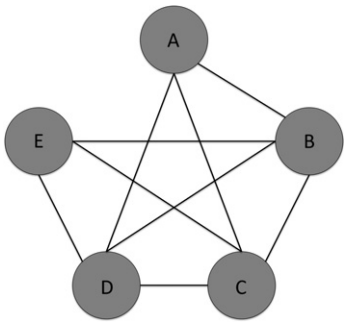


FIGURE 9.1

A network where all pairs of nodes but one are connected.

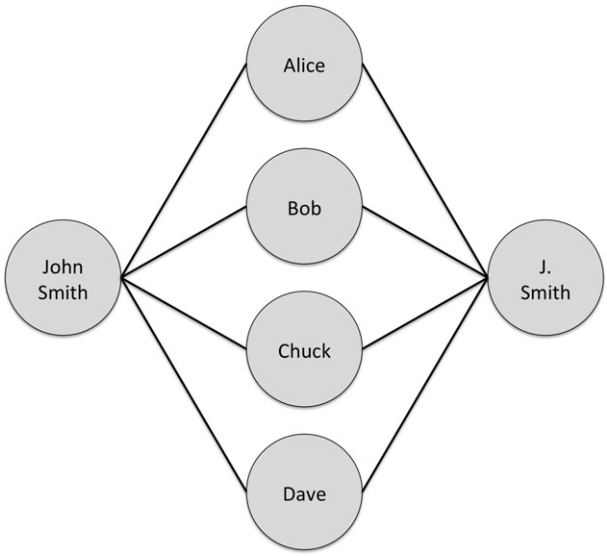


FIGURE 9.2

A network with two nodes, John Smith and J. Smith, who have similar names and acquaintances with no connection to one another. This could suggest that they are actually the same person.

we might conclude that a link should be present. If we consider tie strength, the case can be clearer. For example, in Figure 9.4, the thickness of the edges indicates the tie strength. In Figure 9.4(a), nodes A and E have strong ties with all the other graphs. This forms many forbidden triads, as discussed earlier in the book. It is very rare to have two nodes that share strong ties with another node but have no tie

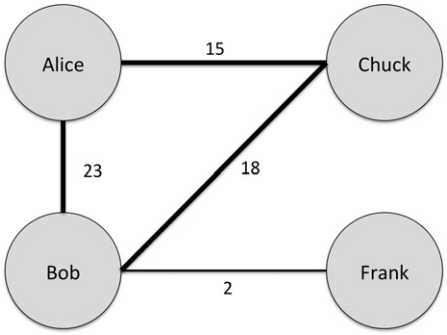


FIGURE 9.3

A network showing the frequency with which Alice, Bob, Chuck, and Frank attend meetings together.

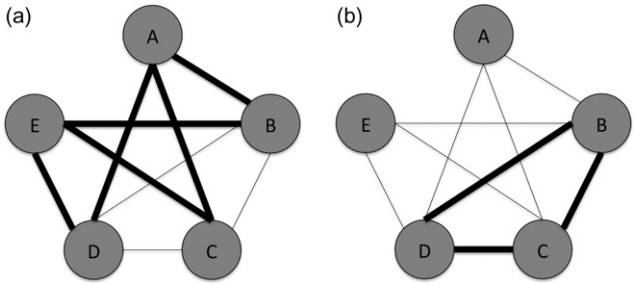


FIGURE 9.4

Two variations of the graph in Figure 9.1 where edge thickness indicates tie strength. In (a), nodes A and E have many shared strong ties, while in (b) they only share weak ties.

with one another. However, in Figure 9.4(b), there are only weak ties between A and E's common neighbors, so it is less likely that they share a tie when compared with graph (a).

These examples provide anecdotes that illustrate what link prediction can do. The next step is to create systematic methods for predicting links.

There are many ways to do link prediction, but all of the algorithms generate a score for each pair of nodes. If we have two nodes, A and B, then the score(A,B) indicates how closely connected A and B are in the graph. After computing the score for every pair of nodes, the algorithm returns a ranked list. The pairs with the highest score are predicted to be the most likely new edges.

As a running example to illustrate how each of the scoring methods works, we will use a simple undirected graph shown in Figure 9.5.

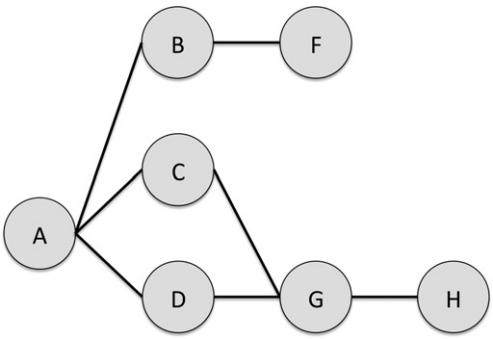


FIGURE 9.5

An example graph with eight nodes.

Mathematical notation

Before looking at the equations for computing scores, we will review some basic mathematical terminology and notation so that we can write the equations concisely. A *set* is a collection of items. In graphs, the neighbors of a node are a set. For example, the neighbors of node A in Figure 9.5 are {B, C, D}. This is a set. Let *Neighbors(A)* indicate the set of A's neighbors. If a set is written with vertical bars on either side, that refers to the size of the set. So |*Neighbors(A)*| means the size of the set of A's neighbors. Since A has three neighbors, |*Neighbors(A)*| = 3. Note that the size of the set of a node's neighbors is equivalent to its degree; that is, *Neighbors(A)*| = *degree(A)*.

Sets can overlap. For example *Neighbors(A)* = {B, C, D} and *Neighbors(G)* = {C, D}. The *intersection* of two sets is the items they have in common. In this example, the intersection of *Neighbors(A)* and *Neighbors(B)* is {C,D} since those nodes are in both sets. The intersection is indicated with the $\cap$ symbol. Thus, to get the neighbors that A and G share in common, we write $Neighbors(A) \cap Neighbors(G)$. The *number* of nodes they have in common is indicated with the vertical bars on either side $|Neighbors(A) \cap Neighbors(G)| = 2$.

The *union* of two sets is the set of all items. The union of *Neighbors(A)* and *Neighbors(G)* is {B,C,D,H}. Note that we do not duplicate nodes C and D.

The symbol Σ is used to indicate that we are taking the sum of values. When working with sets, we might want to take the sum of a value for each item in the set. For example, we may want to add up the degree of each node who is neighbors with node A. To do this, we need to specify that we want each element for a set. In this case *Neighbors(A)* is our set. Then we need to indicate that we want each element in that set. We do this by saying $x \in Neighbors(A)$. That means x represents each item from the set. To show that we are adding these values up,

we put this notation below the Σ. So, to sum the degree of each node who is neighbors with A, we would write:

$$\sum_{x \in Neighbors(A)} degree(x)$$

The important thing to remember with this notation is that by putting the $x \in Neighbors(A)$ underneath the Σ , it means to add up the value after the Σ for each of the items x represents. Although the notation may be a bit complex if you have not seen it before, breaking it down will make it easy to understand.

Computing score

One of the simplest ways to score the similarity or closeness of two nodes is to use the shortest path length between them. Nodes that are close to one another are more likely to create a relationship. This is especially true for nodes that have a mutual friend. However, as the average shortest path length increases, we want the score to decrease because nodes that are far apart (with a high average shortest path length) are less likely to be connected. Thus, we can use the negative value of the shortest path, so closer nodes have higher scores.

$$score(A, B) = -shortestPath(A, B)$$

So for Figure 9.5, the ranked list of scores is as follows:

Pair	Score: -Shortest Path Length
A,F	−2
A,G	−2
B,C	−2
B,D	−2
C,D	−2
C,H	−2
D,H	−2
A,H	−3
B,G	−3
C,F	−3
D,F	−3
B,H	−4
F,G	−4
F,H	−5

Note that since this is an undirected network, each pair appears only once in the list. For example, A,F appears and F,A is not listed since it would have the same value. If the network were directed, node pairs representing an edge in either direction would be listed.

In this first example with the shortest path length, many nodes are tied with a high score of -2. If a simple rule is used to predict that edges will occur between nodes with the highest scores, then all these pairs—(A,F), (A,G), (B,C), (B,D), (C,H), and (D,H)—would have predicted edges between them. Indeed, when using this method, any nodes that have at least one common neighbor will have a predicted edge added between them.

Another way of computing scores that uses more information from the network structure is to count the number of common neighbors between the two nodes in a pair. For the pair (A,B), we can represent this as the intersection of the set of nodes that are neighbors of A and the set of nodes that are neighbors of B.

$$score(A, B) = Neighbors(A) \cap Neighbors(B)$$

The result of this equation is the number of neighbors that the two nodes share. For the graph in Figure 9.5, the results are as follows:

Pair	Score: Common Neighbors
A,G	2
C,D	2
A,F	1
B,C	1
B,D	1
C,H	1
D,H	1
A,H	0
B,G	0
B,H	0
C,F	0
D,F	0
F,G	0
F,H	0

The result here is quite different. Two pairs, (A,G) and (C,D), have the high score. Thus, these would be the only edges predicted when we apply this algorithm.

The number of common neighbors makes social sense, too. The more friends two people have in common, the more likely they are to be introduced to one another.

However, number of common friends is not the whole story. Some people have an abnormally high number of connections in social networks, particularly in social media. For example, celebrities may have millions of friends on Facebook, but the fact that, for example, a popular singer and a politician have many friends in common may not mean much since they are connected to so

many people in the first place. A common statistic called the *Jaccard Index* can account for this.

The Jaccard Index counts the total number of friends in common and divides that by the total number of people who are friends of either node. So, in our simple graph in Figure 9.5, nodes A and G have two friends in common. The total number of nodes who are friends with either A or G is four: nodes B, C, D, and H. Note that we do simply add the number of nodes who are friends with A (3) to the number of nodes who are friends with G (3), because this would count their mutual friends twice (nodes C and D). Instead, we are taking the *union* of their friends.

Note that the size of the union is always the sum of the degrees of the two nodes minus the size of the intersection. For nodes A and G, the sum of their degrees is 6 (3 + 3) and the size of the intersection (number of common friends) is 2, so the size of the union is $6-2=4$, as we counted above. This will be useful later.

Thus, the formula for the Jaccard Index used to compute a score between nodes is:

$$score(A,B) = \frac{|Neighbors(A) \cap Neighbors(B)|}{|Neighbors(A) \cup Neighbors(B)|}$$

For the graph in Figure 9.5, the scores are as follows.

Pair	Score: Jaccard Index
C,D	1
C,H	0.50
D,H	0.50
A,G	0.50
A,F	0.33
B,C	0.33
B,D	0.33
A,H	0
B,G	0
B,H	0
C,F	0
D,F	0
F,G	0
F,H	0

To clarify further, here are the calculations for a few of these pairs. As mentioned earlier, nodes A and G have two common friends and four total nodes in the union of their neighbors. Thus, their score is $2/4 = 0.5$. Nodes C and H also have a score of 0.5, but they share only one friend in common. Since there are only two nodes in the union of their neighbors (nodes A and G), their score is

$1/2 = 0.5$. Node A and F have one friend in common also, but there are three nodes in their union (B, C, and D), so their score is $1/3 = 0.33$. Nodes C and D have two common neighbors (A and G). Since these are the only neighbors of C and D, their score is $2/2 = 1$.

Thus, in this network, we would predict that the next edge appears between nodes C and D.

The value of the Jaccard Index becomes clearer in a big network. Say we have four nodes: Alice, Bob, Chuck, and Dave. Let Alice and Bob be celebrities, each with 1 million friends. Chuck and Dave are average users with 100 friends each. Now say Alice and Bob have 2,000 friends in common while Chuck and Dave have only 20 friends in common.

Although Alice and Bob may seem to be more strongly connected than Chuck and Dave, since they have 100 times more common friends, the Jaccard Index indicates this is not the case. Remember: The size of the union is the sum of the degrees minus the size of the intersection. Thus, the Jaccard scores for these two pairs is as follows:

$$score(Alice, Bob) = \frac{2,000}{(1,000,000 + 1,000,000) - 2,000} = \frac{2,000}{1,998,000} = 0.001$$

$$score(Chuck, Dave) = \frac{20}{(100 + 100) - 20} = \frac{20}{180} = 0.11$$

So, although the number of friends in common is 100 times higher for Alice and Bob, the Jaccard Index is over 100 times higher for Chuck and Dave because they do not have as many total friends. Stepping back from the math, it makes sense that people who have 20 real friends in common are likely to be closer than celebrities who have lots of common "friends" but also far more friends that are not shared.

This example brings up another problem. What if the 20 people Chuck and Dave know in common are also celebrities? That is much less meaningful than if they are other people with a smaller number of friends. Adamic and Adar (2003) came up with a method for dealing with this issue. They look at common friends and assign a score that gives more weight to people who have a few friends.

The formula is as follows:

$$score(A, B) = \sum_{x \in Neighbors(A) \cap Neighbors(B)} 1/\log(|Neighbors(x)|)$$

The formula looks a bit complicated at first, but it is quite simple. For every node who is a neighbor of both A and B (call this node x), we add a value to the total. That value is 1 over the log of the number of neighbors x has. For a node with 100 neighbors, the value would be $1/\log(100) = 1/2 = 0.50$. For a node with

2,000 neighbors, the value would be $1/\log(2,000) = 1/3.3 = 0.30$. As the number of neighbors increases, the value decreases. A node with 1,000,000 neighbors would only have a value of 0.17.

The values for our sample network using the Adamic/Adar method is as follows:

Pair	Score: Adamic/Adar
A,G	6.64
C,D	4.19
A,F	3.32
B,C	2.10
B,D	2.10
C,H	2.10
D,H	2.10
A,H	0.00
B,G	0.00
B,H	0.00
C,F	0.00
D,F	0.00
F,G	0.00
F,H	0.00

The clear winner here is (A,G). This method predicts that the next link to be added is between these nodes. Note that in the Jaccard measure, the pair (C,D) came out ahead. They are lower on the list here because their neighbors, A and G, both have the highest degrees in the network and thus they do not count as strongly.

While it is different from the Jaccard Index rankings, the ranking here is the same as when we used the number of common neighbors. This is because, in our sample network in Figure 9.5, all the nodes have a small degree. Thus, the method here will not have a large impact on the scores. However, when there are large differences in the degrees of nodes, as is expected in most networks since the degree follows a power law distribution, there will be larger effects from using this approach.

One final technique for predicting links is to consider *preferential attachment*. This network principle states that nodes with a high degree are more likely to gain new links. Popular nodes are more likely to gain new friends than less popular nodes. When predicting edges, preferential attachments suggest that nodes with high degree are more likely to gain new edges. The formula for this scoring method is relatively simple:

$$score(A, B) = |Neighbors(A)|^*|Neighbors(B)| = degree(A)^*degree(B)$$

For the example in Figure 9.5, the scores are as follows:

Pair	Score: Preferential Attachment
A,G	9
B,G	6
B,C	4
B,D	4
C,D	4
A,F	3
A,H	3
F,G	3
B,H	2
C,F	2
C,H	2
D,F	2
D,H	2
F,H	1

With this measure, we would predict that the next edge to appear will be between nodes A and G.

Advanced link prediction techniques

The examples covered so far are relatively straightforward link prediction techniques, and there are many ways to make the approach more sophisticated. One could begin by combining the measures above. For example, we could take the average ranking of each node pair from each measure and rank by that value. The result would be a ranking that considers all the factors described above.

There are also probabilistic models for link prediction that are very successful. These often rely on a technique called Markov Networks . Some approaches consider nodes' attributes in addition to network structure. They can also work with weighted and directed graphs. Machine learning has also been effective when applied to this problem.

While these techniques are beyond the scope of this book, many references can be found online. Good overviews are also provided in Getoor and Diehl (2005) and Liben-Nowell and Kleinberg (2007).

Entity resolution

Entity resolution is a technique that tries to identify nodes that represent the same entity and then to merge them together. For example, in Figure 9.2, the two nodes

"John Smith" and "J. Smith" may represent the same person. How do we determine if they are the same or not?

Just as there are many techniques for doing link prediction, there are a number of methods for entity resolution. Most of them involve looking at the data about the nodes, including their attributes and relationships.

Table 9.1 contains sample data for four people. Before getting into network connections, we can look at this information alone.

The simplest approach to merge duplicated nodes is used when we have unique identifiers for each node. For example, each person in the United States has a unique Social Security Number (SSN). Each person has only one SSN, and each SSN is used for only one person at a time. In Table 9.1, nodes "J Smith" and "John Smith" have the same SSN, so we know they must represent the same person. If their SSNs were different, we would know they are definitely not the same person.

Other attributes can allow us to make similarly definitive conclusions, but not as often. For example, the birthday of a person should be consistent. In this case, "John Smith" and "JA Smith" have the same address, but their birthdays are different. Thus, we can conclude either that they are not the same person or that there is an error in the data. We will assume the data is correct in these examples to more easily illustrate our points.

Not all attributes need to match. For example, "J Smith" and "John Smith" have different addresses. People move or have work and home addresses, so the mismatch on that point does not indicate that the nodes are or are not the same person.

Similarly, first names do not need to match. People may use nicknames, they may go by their first and middle names in different contexts, and their initials may be used. Similarity in names can suggest that two people are the same, but that is not totally conclusive.

For example, in Table 9.2 we have "J Smith" and "Robert Smith." Their names are similar in that they have the same last name. The differences in the first name could be because the same person is using his first initial in some cases and his middle name in others. Or they simply may be different people. If we look at the other data, we see that they have the same address and birthday. Those shared attributes provide evidence that they may indeed be the same person.

Table 9.1 Sample Personal Data for Four People

First Name	Last Name	Address	Birthday	SSN
J	Smith	123 Main St	1/6/68	123-12-1234
John	Smith	54 Elm St	January 1968	123-12-1234
Robert	Smith	123 Main St	1/6/1968	
JA	Smith	54 Elm St	March 1968	

Table 9.2 Values for each Example Node Pair and the Associated Similarity Measures

	Node Pair		
	A,J	**B,D**	**E,I**
Common Neighbors	3	0	1
Jaccard Index	0.38	0	0.33
Adamic/Adar	9.97	0	1.43
Preferential Attachment	25	1	4

There is uncertainty at this point. How can we deal with that? We use a similar approach to that from link prediction: scoring. For each pair of nodes, we can compute a score that represents the likelihood that they are the same node. Then, we can set a threshold value. Any pair of nodes with a score above that threshold will be merged, and any nodes below the threshold will not be merged.

Scoring techniques

There are many approaches for creating scores, and these can become quite sophisticated, using machine learning algorithms and data mining. In this chapter, we will present one of the simpler approaches that you could apply on your own.

To create a score for a pair of nodes, we will consider similarity on a set of attributes. Those can include data such as that in Table 9.2 and similarity in the network structure. For each pair of nodes, we will know if their values match or not on a given attribute. For example, in Table 9.2, "J Smith" and "John Smith" have the same SSN. Thus, we can record a score of 1 for that attribute indicating a match. However, their first names do not match. In that case, we record a score of 0.

As discussed earlier, however, matching on some attributes is more important than on others. Two different people may have the same name, but a match on the SSN is much more definitive. Thus, we would want to give more weight to a SSN match than we do to a name match. Similarly, we want a weight for when nodes do not match. For example, if nodes do not match on their SSN, we want to subtract a lot of value from the score since that is a strong indicator that the nodes represent different people.

To create a score for a pair of nodes, we will determine that they match on a given attribute and we will have a weight for each attribute. Then, we add the positive weights for each attribute where the nodes match (i.e., receive a score of 1), and subtract the negative weights when they do not match (i.e., receive a score of 0).

Then, we are left to find a method for computing the weights for each attribute.

We want weights to be higher for attributes that are more definitive, like the SSN, and lower for attributes that are more commonly shared, like the month of birth. There is a common method for computing these weights for the entity resolution. This is done with values called *u* and *m* probabilities. The *u* probability is that two nodes will match on an attribute by chance. For example, the probability that two nodes have the same birth month is 1/12. Thus the *u* probability for birth month equals 1/12 or 0.083. The probability of two nodes having the same last name is more complex to compute because the probability varies based on the last name itself. For example, "Smith" is the most common last name in the United States, representing about 1% of all citizens' last names. Thus, the *u* probability for matching on the last name "Smith" is 0.01. However, for an uncommon last name in the United States, like "Himmelblau," which is used by only 0.00004% of the population, the *u* probability would be 0.0000004. When computing *u* probabilities, we can have a single value, like for birth month, or a set of values for each possible attribute value, like last name.

The *m* probability is the probability that two nodes that represent the same person will have the same value. Often we expect this value will be 1. For example, two nodes that are the same should have the same birthday, gender, SSN, and so on. However, the *m* value is not always 1. In some cases, like address or phone number, two nodes may indeed represent the same person but have different values. For example, one node could have personal/home information, and the other could have work information. Also, there may be missing attribute data. For example, in Table 9.2, several nodes are missing SSNs. Thus, they could represent the same person, but if one has an SSN and the other does not, the values will not match.

Setting the *m* probabilities will depend on the data you have. For our data, we could say the *m* probability for SSN is 0.95 (assuming there is more data than what is shown in Table 9.2 and we know how good it is), the *m* probability for address is 0.6, and the *m* probability for birth month is 0.98.

Once we have the *u* and *m* probabilities, we need to turn them into weights. There will actually be two weights for each attribute. The first is how much weight we add to the score if there is a match, and the second is how much weight we subtract from the score if there is no match. The common formulas are as follows.

For a match:

$$w = \ln(m/u)/\ln(2)$$

For a nonmatch:

$$w = \ln\left(\frac{1-m}{1-u}\right)/\ln(2)$$

Using the values we discovered above, the weight for a match on birth month would be:

$$\ln(0.98/0.083)/\ln(2) = 2.469/0.693 = 3.56$$

The weight for a no-match on birth month would be:

$$\ln\left(\frac{1-0.98}{1-0.083}\right)/\ln(2) = \ln\left(\frac{0.02}{0.917}\right)/\ln(2) = -3.825/0.693 = -5.520$$

We would perform this calculation for every attribute in the table. Then, we would check for matches and add the appropriate weights for a match or non-match to compute a final score.

Incorporating network data

The scoring above works with fixed attributes for a set of nodes. It does not look at the network structure at all, and that can be very important. For example, in Figure 9.2, John Smith and J Smith share many friends in common, but they are not connected to one another. If John and J are different people, we would probably expect them to know one another since they have so many common acquaintances.

Relational data is useful for enhancing the attribute-based similarity discussed above. Consider a more sophisticated graph than the one in Figure 9.2.

We could compute similarities between all pairs of nodes in the network. For simplicity, we will consider three pairs of nodes as examples: A and J, B and D, and E and I.

Before considering any formulas and just by observing the network, some features emerge. Nodes A and J both share several common neighbors, and they also have the highest degrees in the network. Nodes E and I have a common neighbor but have much lower degrees. Nodes B and D are far apart in the network. Without any mathematical work, we might consider that A and J seem most similar and B and D seem most distant.

To quantify how similar these nodes are to one another structurally, we want to examine their egocentric networks and compare them. Specifically, we want to compare the neighbors of one node to the neighbors of the other. This is exactly the same comparison we made when performing link prediction above. Thus, we can use many of the same scoring mechanisms from link prediction to quantify how similar a pair of nodes are to one another.

For entity resolution, the number of common neighbors, the Jaccard Index, the Adamic/Adar method, and preferential attachment all compared the neighbors of one node with those of another. We will use those same measures here. Table 9.2 gives the values for each measure to each pair of nodes in our example.

To review, the common neighbors simply counts how many nodes are neighbors of both nodes in the pair. Nodes A and J have three common

neighbors: nodes E, F, and G. Nodes B and D have no common neighbors. Nodes E and I have one node, J, as a common neighbor.

The Jaccard Index divides the size of the intersection by the size of the union. Nodes A and J have three nodes in their intersection (E, F, and G), and eight nodes in the union (B, C, E, F, G, H, I, and K). Thus, the Jaccard Index is $3/8 = 0.38$. Since nodes B and D have no neighbors in common, the Jaccard Index is 0. For nodes E and I, they have one node in common and three nodes in their union (A, J, and K). This yields a Jaccard Index of $1/3 = 0.33$.

The Adamic/Adar method sums up the inverse log of each neighbor's degree, giving more weight to nodes with fewer edges. For nodes A and J, this results in the following sum:

$$1/\log(degree(E)) + 1/\log(degree(F)) + 1/\log(degree(G)) =$$
$$1/\log(2) + 1/\log(2) + 1/\log(2) = 3/\log(2) = 3/0.301 = 9.97$$

For nodes E and I, they only have node J in common. That gives a simpler result:

$$1/\log(5) = 1.43$$

Finally, preferential attachment is the product of the degree of the two nodes being considered. For nodes A and J, that product is $5*5 = 25$. For nodes B and D, the product is $1*1 = 1$. This is the only measure that is nonzero for this pair. Nodes E and I have a product of $2*2 = 4$.

The results from these similarity measures can be used in addition to attribute data. For example, if two nodes are very similar in their attribute data but have very little similarity in the network, we can reduce the similarity score. A high similarity on the network may make up for lower similarity in attribute data as well. Network and attribute data can be considered as separate steps, or the network data score can receive its own weight for use in the sum above.

More sophisticated entity resolution

As with link prediction, there are many more sophisticated methods for doing entity resolution that are beyond the scope of this book. These use machine learning techniques, Bayesian networks, and statistical models. A good overview is available in Brizan and Tansel (2006).

However, there are some ways to iterate on even the relatively simple methods introduced here. One approach is to allow for partial matches. Returning to the "John Smith" and "J Smith" example, while their first names are not an exact match, they are close. Since "J" is the correct first initial for "John," we may label this a partial match. Then, instead of adding the weights for items that match, we can add part of the weight for a partial match. For example, if we say "J" is a 0.3 match for "John," then we could add 0.3 times the weight for a name match to the score. We would also have the option of subtracting 0.7 times the non-match score.

In this approach, the formula for an attribute that is a partial match with value p, we would add the following to the score:

$$p^* w_{match} + (1 - p) w_{non-match}$$

Say the weight for a matching first name is 5.5 and the weight for a non-matching first name is −3.2. If we did not give any credit for a partial match, then we would simply subtract 3.2 from the score. But if there is a 0.3 match on the first name, then the score becomes:

$$0.3^*5.5 + 0.7^* - 3.2 = 1.65 - 2.24 = -0.59$$

This partial match allows us to give much more credit to the pair, subtracting only 0.59 instead of 3.2.

A second, more sophisticated step is that we can do repeated iterations of entity resolution. For example, say we merge nodes A and J in the graph in Figure 9.6. After they are merged, the new graph that results is shown in Figure 9.7.

Now, node H has many similar neighbors with the merged node A/J. In fact, if we recompute the measures of similarity on the network, the scores for nodes H and A/J are as high or higher than they were for A and J in the previous round. This is shown in Table 9.3.

The network data suggests a lot of similarity between nodes H and A/J. If the attribute data supported a decision to merge node H with node A/J, we would produce a second new graph, shown in Figure 9.8.

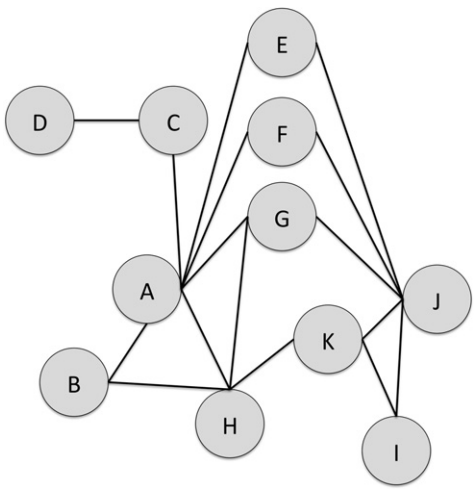


FIGURE 9.6

A graph in which we will consider whether or not to merge nodes. The examples will consider merging A and J, B and D, and E and I.

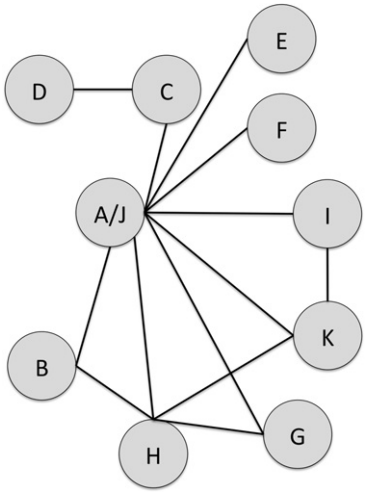


FIGURE 9.7

The network from Figure 9.6 after nodes A and J are merged.

Table 9.3 Measures of Network Similarity for Nodes on the Merged Network Shown in Figure 9.7

	Node Pair			E,I
	Previous A/J	**A/J, H**	**B,D**	
Common Neighbors	3	3	0	1
Jaccard Index	0.38	0.43	0.00	0.50
Adamic/Adar	9.97	9.97	0.00	1.11
Preferential Attachment	25	32	1	2

Note that the values for the pair E and I have also changed because of the merger in the network.

Attribute data should be considered in this iterative process as well, and similarities between other nodes in that respect may lead to further merges.

Link prediction: Case study—Friend recommendation

Link prediction is an important and widely studied problem, but what are the applications of a good link prediction algorithm? There are many, and one case particularly relevant to the topics in this book is for friend recommendation.

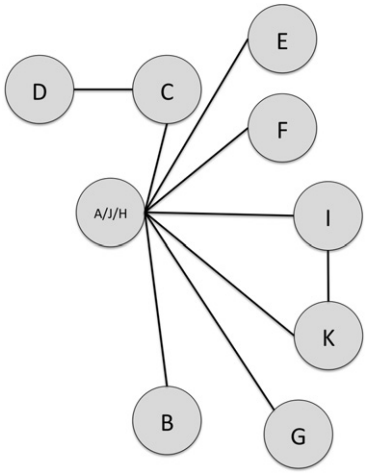


FIGURE 9.8

The network from Figures 9.6 and 9.7 with nodes A, J, and H all merged.

FIGURE 9.9

A suggestion about people to follow made by Twitter.

Many social networking and social media websites have a feature that recommends friends. For example, Figure 9.9 shows Twitter's "Who to follow" recommendation.

How does a system go about recommending people to friend or follow? There are many techniques, and link prediction is one way to do it.

Recall that link prediction, as described earlier, considers all unconnected pairs of nodes in the network and generates a score for each. Those scores can be

used to add the top-scoring link to the graph, or they can be considered a ranked list of potential edges to add. For friend recommendation, we do not necessarily want to consider all edges in the graph, but rather all possible edges for a specific user.

When that user logs in, the system can compute a score for each pair comprising the user and every other node in the network. Then, the pairs can be sorted from highest to lowest score, and the other node in the top-scoring edges becomes a recommended friend.

For large networks, however, computing scores for every pair of nodes can be computationally expensive and take a long time. For example, Facebook has over a billion users. Running 1 billion calculations takes a long time, especially if the system needs to get the friend list for every person. Fortunately, the process can be optimized. Recall that for most of the link prediction techniques, the nodes needed to share at least one common neighbor. If the system uses this as a limit, then the only nodes that need to be considered as candidate friends for the user are the users' friends' friends. That greatly cuts back on the number of possible pairs to score, making the computation much faster.

Note that link prediction results are not necessarily the only thing to consider when recommending friends. Looking at similarity of node attributes can add valuable information. While the interesting attributes will be different from those in entity resolution, the techniques for using them may be similar. For example, when recommending friends, we might look for people with matches on interests, educational background, favorite sports teams, and so on. We can create weights for matches on each of those attributes and use them in a score, just as we used weights on matching personal information to conduct entity resolution. Combining these attribute-based insights with the link prediction results will often lead to better friend suggestions.

Entity resolution: Case study—Finding duplicate accounts

When people sell things online on sites such as eBay, Etsy, or Amazon, the transaction requires that the buyers trust them. Even with insurance and seller protection systems in place, few buyers want to go through the hassle of receiving a bad item or no item and then filling out forms and dealing with a system's bureaucracy to receive a refund. Thus, the seller's reputation is extremely important for the transaction to go well.

When sellers develop a poor reputation, a common "solution" is to open another account. Having no reputation is often better than having a poor one. In more sinister cases, some sellers will develop good reputations in many accounts by selling small items, leveraging that reputation to sell a few big items at which point they defraud the buyers, absconding with the money and closing the now-worthless account.

To protect buyers, companies that host online sales want to ensure that people are not maintaining multiple accounts without an obvious link between them. Knowing which accounts belong to the same person allows the company to track the good and bad actions of each unique person and to have the power to suspend *all* the accounts if the seller does something very bad on one of them, or if the sum of bad behavior across the accounts crosses some threshold.

When sites have millions of sellers, as many do, how can a company track which accounts actually belong to the same person? Entity resolution works well for this task. User attributes, like financial information and addresses, are often very distinctive and can help identify the accounts' owner. Network information can also be included, especially when the accounts are linked to the same products or customers.

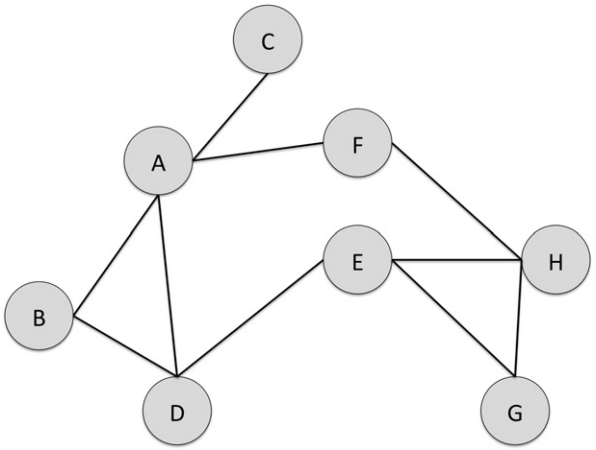
Conclusion

Link prediction and entity resolution are two ways to identify missing information in networks. Link prediction helps identify edges that are likely to appear in the future, if they do not exist already. Entity resolution uses attributes and network structure data to link nodes that represent the same individual.

These techniques take advantage of many network features covered earlier in this text, including degree, clustering, and path lengths. This chapter introduced some of the basic methods for both tasks. As described above, many more sophisticated computational approaches to both link prediction and entity resolution exist, and those will make excellent further reading for computer scientists interested in this topic.

The results can be applied in many areas. Two short case studies discussed how link prediction can be used to recommend connections in social media and how entity resolution is useful for identifying duplicate accounts belonging to the same person. Link prediction is also particularly useful for network forecasting. Knowing which people in an organization are likely to connect can be used in many ways. Within companies, for example, this could be leveraged to make introductions and get collaborations moving faster. Within a criminal or terrorist organization, the predicted links could provide interesting intelligence about how the group will evolve. Entity resolution also has many other applications in social media and online. It is often applied to Census records, where data about people in multiple locations should be connected. It has similar anticrime and antiterrorism applications, linking aliases with true identities. Other applications include merging duplicate products in online shopping, merging duplicate web search results, and detecting spam.

Exercises



1. In the graph above, there are 28 pairs of nodes. Ten of those pairs already have edges between then (e.g., A and B, E and H). The remaining pairs have no direct connection. For each of the indirectly connected pairs, complete the following table with scores for the indicated link prediction formula.

Pair	Shortest Path	Common Neighbors	Jaccard Index	Adamic/ Adar	Preferential Attachment
A,E					
A,G					
A,H					
B,C					
B,E					
B,F					
B,G					
B,H					
C,D					
C,E					
C,F					
C,G					
C,H					
D,F					
D,G					
D,H					
E,F					
F,G					

2. For each of the formulas in the table for Exercise 1, list the pair of nodes or pair between which you would add an edge based on the scores. Assume only the top-rated pair is selected.
3. Are there any patterns that emerge in your analysis of the graph in exercises 1 and 2? Are there certain pairs that frequently receive high or low ratings? Can you explain that by analyzing their position in the graph structure?
4. Compute the *u* probability for the following attributes:
   a. Gender
   b. Marital status. Assume the options are married, single, divorced, and widowed and for simplicity, assume there are the same number of people in each category, like with gender or birth month.
5. On the website companion for the book, you will find a file with the 2010 U.S. Census data showing the population data for each zip code in Washington, D.C. The total population of D,C. in that dataset is 601,723. Compute the *u* probability for the following zip codes based on that dataset:
   a. 20010
   b. 20045
   c. 20535
   d. 20002
6. If we know our dataset is 100% correct with no errors and no missing data, what is the *m* probability for matching nodes to have the same birthdate?
7. Again, assuming the data is 100% correct with no errors or missing data, is the *m* probability for phone number 1? Why or why not?
8. Assume the following table describes the nodes in the graph above.

Node	First Name	Last Name	SSN	City	State	Marital Status
A	John	Doe	123-23-1324	Chicago	IL	Married
B	Jane	Doe	234-32-4321	Chicago	IL	Married
C	Robert	Donovan				Divorced
D	John	Smith	132-13-1321	Washington	DC	Single
E	William	Smith		Washington	DC	Single
F	Jeannette	D.	234-32-4321	Madison	WI	Married
G	Jeannette	Doe		Seattle	WA	Single
H	JB	Doe	123-23-1324	Madison	WI	Widowed

Using the same list of node pairs from exercise 1, compute the scores and fill in the table below using the following match and nonmatch scores. Use only matches or nonmatches, not partial matches in your scoring. Show your work.

Score	First Name	Last Name	SSN	City	State	Marital Status
Match	2.1	3.4	6.1	4.9	1.6	1.8
Nonmatch	−1.3	−4.8	−3.3	−2.7	−1.9	−2

Pair	Score
A,E	
A,G	
A,H	
B,C	
B,E	
B,F	
B,G	
B,H	
C,D	
C,E	
C,F	
C,G	
C,H	
D,F	
D,G	
D,H	
E,F	
F,G	

9. Repeat exercise 8 but allow for partial matches. You can assign your own values between 0 and 1 for a partial match. For example, if one node's first name is "Michael" and another node has a first name listed as "M," you may decide to use a value of 0.3 as a partial match because the initial "M" could represent "Michael."

List each partial match, your value for it, and your reasoning for the value. Then, recompute the score for each pair of nodes listed above. Show your work.

10. For the following node pairs, compute their similarity for entity resolution using the specified methods.

Pair	Common Neighbors	Jaccard Index	Adamic/ Adar	Preferential Attachment
A,E				
A,B				
A,H				
B,C				
B,E				
C,F				
D,E				
D,H				
E,F				
F,H				

11. Give a weight of 7.3 to the Jaccard Index as you computed it in exercise 10. For the following pairs, give a new score that incorporates the Jaccard Index in addition to the attribute data. Do this by adding the Jaccard Index times its weight to the existing scores. Do this for the full matches (from exercise 8) and partial matches (from exercise 9).

Pair	Full Match Score	Partial Match Score
A,E		
A,H		
B,C		
B,E		
C,F		
D,H		
E,F		

12. Set the similarity threshold equal to 4.0.

- **a.** Which nodes will be merged based on your calculations in exercise 8?
- **b.** Which nodes will be merged based on your calculations in exercise 9?
- **c.** Which nodes will be merged based on your calculations in exercise 11?
- **d.** How does incorporating the network similarity data change the results of the entity resolution decision?

13. The conclusion listed several applications of entity resolution and link prediction. Pretend you are running a social network. Come up with your own new applications—one for link prediction and one for entity resolution. Describe the problem you would solve and how you would use the techniques presented in this chapter.

This page intentionally left blank

CHAPTER

Propagation in Networks 10

Social networks allow all types of things to spread from person to person. Diseases, viral videos, fads, rumors, and many types of information all propagate from one person to another in networks. Interestingly, many of these things can be modeled in similar ways. Thus, understanding how diseases spread through networks also provides an understanding of how fads, rumors and many other things spread.

This chapter will introduce some fundamental ways of modeling and understanding propagation in networks, and show how they apply in some specific case studies.

Epidemic models

Diseases and their spread have been studied extensively. Many factors will impact whether persons who are exposed to a disease catch it; their age, nutritional state, overall health, natural immunity, length of their contact with the disease carrier, and the environment are a few examples. However, it is impossible to know every factor for every person in the population; thus, it is necessary to make simplifying assumptions to model the process.

One way to do this is with *compartmental models*. This categorizes people according to their state with respect to the disease. There are three major categories, each represented by a letter:

- S (Susceptible). These are people who have not yet contracted the disease, but who are susceptible to catching it.
- I (Infected). These people have caught the disease and are actively infected and contagious.
- R (Recovered). People in this state have recovered from the disease and are no longer contagious or susceptible to reinfection.

Combining the letters for each state describes the disease model. Common models include SI, SIR, SIRS, and SIS.

For example, an SI disease is one where a person is susceptible, then can become infected, but once infected, never recovers. HIV is a classic example of this type of disease. An SIR model describes a disease where a person is susceptible, becomes infected, recovers, and from then on is immune to the disease. The

Table 10.1 A List of Four Disease Models and Example Diseases for Each

SI	SIR	SIS	SIRS
HIV	Chicken Pox	Strep throat	Whooping Cough
Herpes	Mononucleosis	Common Cold	Syphilis
	Mumps		Chlamydia
	Rubella		Salmonella
	Measles		
	Polio		

chicken pox and mononucleosis are two common examples of these. Once a person becomes sick and recovers from one of those diseases, they can (generally) never catch them again.

An extension of the SIR model is the SIRS. In SIRS, a person is susceptible, becomes sick, recovers, and enjoys a period of immunity before becoming susceptible again. Whooping cough, a bacterial disease that causes a weeks-long severe cough, fits this model. Once infected, a person is sick for a couple months, and after recovering the person is temporarily immune before becoming susceptible again.

Table 10.1 lists several disease models and example diseases for each.

Threshold models

In the compartmental models discussed above, a person is either susceptible to a disease or not; there is no consideration given to the level of susceptibility or to what makes a person more or less likely to catch a disease. A threshold model considers how many infected individuals a person must be exposed to before becoming infected.

This is called a *k-threshold model*, and k represents the number of neighbors who must be infected for a node to catch the disease. If someone can become infected from only one neighbor, this is a 1-threshold model. If three neighbors must be sick for the disease to be passed, it is a 3-threshold model.

Consider the simple network shown in Figure 10.1 as an example. There are 16 nodes connected in a grid pattern. Node 5, indicated in black, starts off as infected. In a 1-threshold model, nodes 1, 6, and 9 could be infected by their sick neighbor, node 5. In a 2-threshold model, no nodes can be infected because only node 5 is sick, and thus no node has 2 sick neighbors. This is true for any $k > 1$ in this graph since only one node is sick and thus no node will ever have more than one sick neighbor.

Infections spread in steps; there is an initially infected node, then its neighbors are infected, then those nodes' neighbors, and so on. More formally, these stages can be treated as time steps. The network begins at time 0, written $t = 0$. Then, at

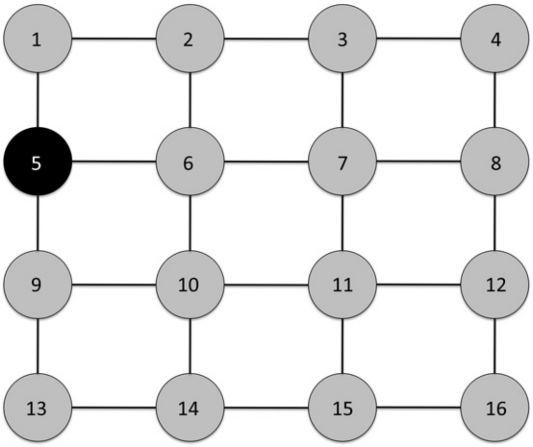


FIGURE 10.1

A simple network with 16 nodes connected in a grid pattern only one of which is infected.

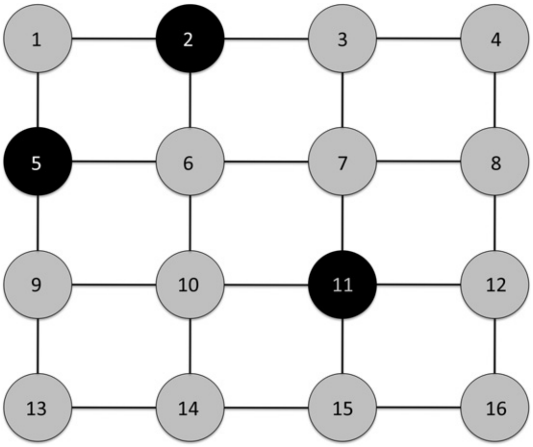


FIGURE 10.2

A grid network with three infected nodes.

the next step ($t = 1$), the first set of neighbors are infected. At time $t = 2$, the next set are infected, and so on.

Now consider Figure 10.2, where nodes 2, 5, and 11 are infected at time $t = 0$. In a 1-threshold model, all the neighbors of these nodes will be infected at step $t = 1$ - nodes 1, 3, 6, 7, 9, 10, 12, and 15. At time $t = 2$, all the remaining nodes will be infected because their neighbors will be sick.

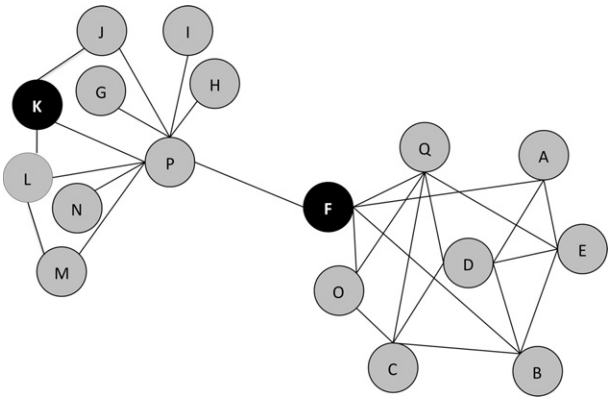


FIGURE 10.3

A more complex network with two infected nodes: F and K.

In a 2-threshold model, things are different. Nodes with two sick neighbors will be infected. At first, that will only be nodes 1 and 6 (infected by 2 and 5). Then, at time $t = 1$, nodes 7 and 10 will be infected by the combination of 6 and 11. At time $t = 2$, nodes 3 and 9 will be infected by 2 and 7, and 5 and 10, respectively. Then, the infection stops. None of the remaining nodes are connected to two infected nodes, so the disease cannot progress.

The same strategy can be applied to more complex networks. Figure 10.3 shows a network with two nodes, F and K, infected. In a 1-threshold model, all the direct neighbors of nodes F or K will become infected at time $t = 1$: O, B, A, Q, L, P, and J. At time $t = 2$, all of those nodes' neighbors will be infected. Since P is sick at this time step, all the nodes in the upper left cluster will be infected. Furthermore, the remaining susceptible nodes in the lower right (C, D, and E) are all connected to sick nodes when $t = 2$, so they too will become infected.

In a 2-threshold model, however, the story is very different. Node P is the only node that is neighbors with F and K, so it becomes infected. Then, at time $t = 1$, in the upper right, nodes L and J will become infected, since they are neighbors of both K and P. Finally, at time $t = 2$, node M is infected by L and P. After that, the infection stops. No nodes in the lower right cluster become infected because F is the only infected contact they have. None will ever be connected to two sick nodes.

The firefighter problem

In the example with Figure 10.3, several nodes seem very important. Node P is a hub, connected to nine other nodes. If P becomes infected, it can pass the

infection on to a lot more people. Depending on the value of k, not all of node P's neighbors will necessarily become infected, but the possibility of infection is higher when node P is sick. If we wanted to stop the infection from spreading, eliminating the possibility of node P becoming sick would have a big impact.

This line of thinking leads to important questions for analyzing propagation in networks. Can certain nodes be "vaccinated" or removed from the network so that they are no longer susceptible to infection and thus will never be able to pass it on? Conversely, if we want to spread something, like a viral video, are there nodes whose network position allows them to reach a wider audience than other nodes?

The *Firefighter Problem* is one way to think about this. In the Firefighter Problem, think of nodes as trees in a forest and as the spreading disease being a fire. Trees can catch on fire if neighboring trees are on fire, unless a firefighter is there to prevent it. Using a k-threshold model, a tree will catch on fire if k of its neighbors are on fire, unless there is a firefighter. At time $t = 0$, an initial set of fires is present in the network. Then, at time $t = 1$, we can place some number (call it n) of firefighters onto trees in the network. At time $t = 2$, the fire will spread to susceptible trees that are not protected by firefighters. At time $t = 3$, we can place n more firefighters, and at $t = 4$, the fire spreads again. The problem progresses in alternating turns until the fire is stopped or all nodes are on fire (i.e., infected).

The point of the Firefighter Problem is to think about how to stop an infection or to make it spread further. It is a theoretical problem studied by graph theorists on very complex graphs because it is mathematically interesting, but it is also practical for people studying the spread of information or vaccination procedures in social networks.

As an example, consider the simple grid network from Figure 10.1 again. As parameters for the exercise, use a 1-threshold model for infection and allow two firefighters to be placed at each turn. At time $t = 0$, node 5 is infected. Then, we can place two firefighters. They can go anywhere, but we will put them at nodes 1 and 6. Then, at time $t = 2$, the disease spreads. Now nodes 1 and 6 are protected, but node 9 is adjacent to node 5 and is susceptible, so it becomes infected. At time $t = 3$, we can place two more firefighters. Nodes 10 and 13 are adjacent to the newly infected 9, and they are the only two nodes we know can get infected at time $t = 4$ if they are not protected, so the firefighters are placed there. That stops the fire from reaching any new nodes, so the fire stops. This is illustrated in Figure 10.4.

The exercise can also be repeated on more complex networks. Again, the disease will follow a 1-threshold model, but this time, only one firefighter can be placed in each turn. Consider the graphs in Figure 10.5, which have the same structure as the network in Figure 10.3.

Nodes A, F, and K begin infected. At time $t = 1$, we can place one firefighter. As mentioned above, node P is a hub, and it can pass the disease on to many neighbors. We will protect node P in the first turn. Then, at time $t = 2$, the disease spreads to neighbors of the infected nodes: B, D, E, J, L, Q, and O. At time $t = 3$,

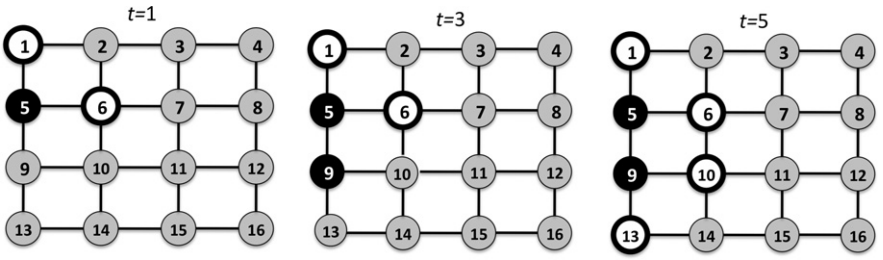


FIGURE 10.4

The placement of firefighters (white circles with black outlines) and progression of the infection (black nodes) at time steps 1–3. After time $t = 3$, there are no susceptible nodes adjacent to the infected nodes, so the disease stops spreading.

we can place another firefighter. Only two nodes are adjacent to infected nodes at this point: nodes M and C. We can choose either to block, and the other one will be infected at time $t = 4$, after which no nodes will be adjacent to an infected node and the infection stops.

These seem like good choices, and in the example network in Figure 10.5, protecting node P was an obvious choice, but how is "success" measured in the Firefighter Problem? One simple way is to count the number of nodes that remain uninfected. After time $t = 4$, node P is uninfected because we protected it, as is either node M or C, which we chose to protect at time $t = 3$. Furthermore, nodes G, H, I, and N all remain uninfected. Thus, a total of six nodes were uninfected.

Compare this to a different choice. Say, instead of protecting node P at time $t = 1$, we chose to protect node L? This is illustrated in Figure 10.6. Then at time $t = 2$, the same nodes in the lower right (B, D, E, O, and Q) are infected, along with nodes J and P in the upper left. We place another firefighter at time $t = 3$. Examining the network reveals that all the remaining nodes will be infected at time $t = 4$, so our choice at this step will only protect that single node—it will not stop further spread. Let's say we choose node M. Then all the remaining nodes are infected at time $t = 4$: C, G, H, I, and N. With this strategy, only two nodes—L and M—were protected. That is far fewer than the six nodes protected when we chose to place the firefighter on node P.

This comparison is an important way to judge which nodes are most effective to protect or, on the flip side, which nodes can spread a disease or information most effectively.

Stochastic models

The compartmental models and threshold models discussed above are *deterministic* models; they assume that any susceptible individual who is exposed to an

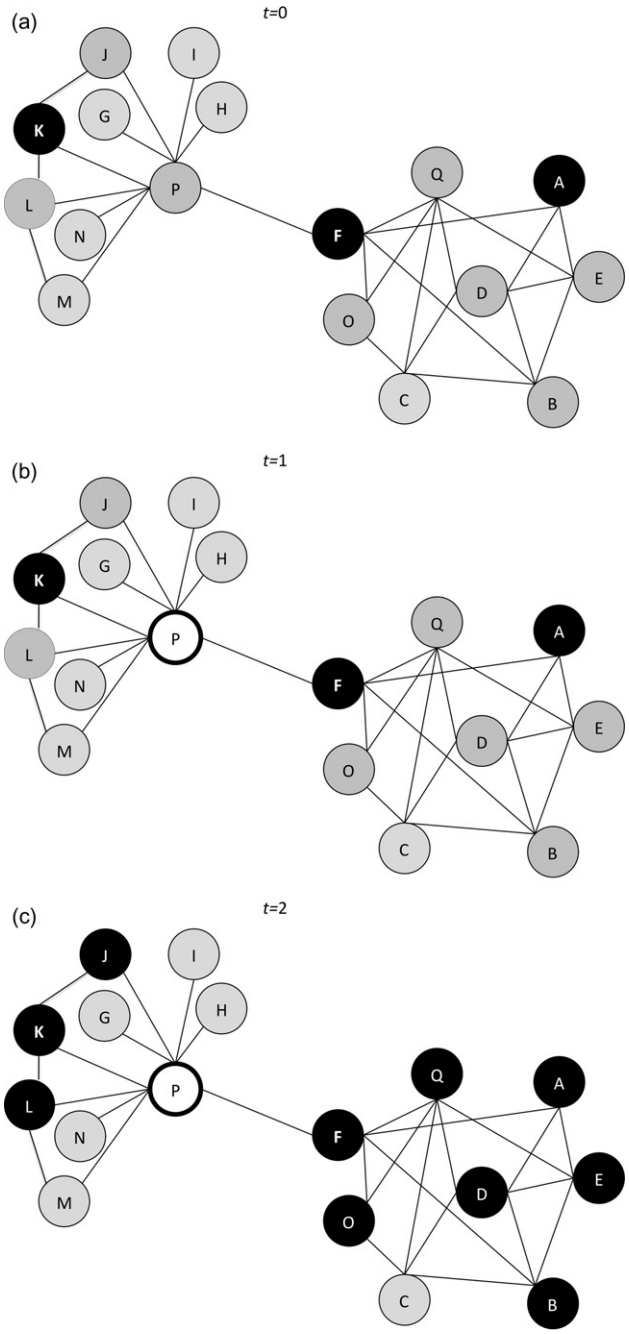


FIGURE 10.5

The spread of infection and placement of firefighters in a more complex network.

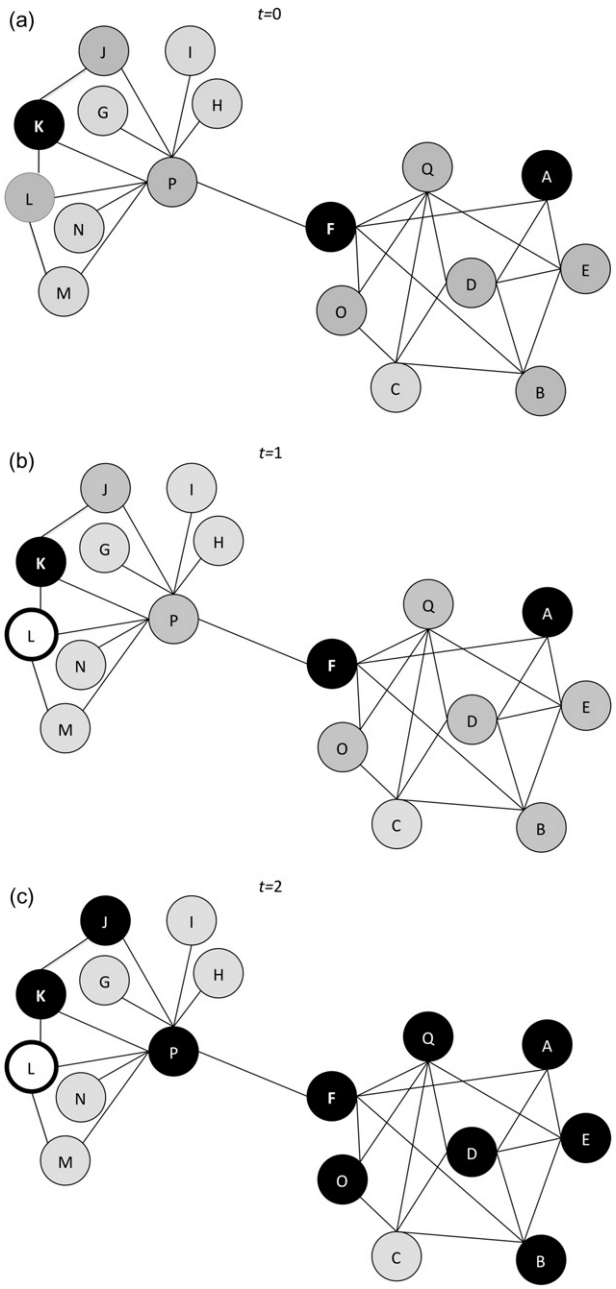


FIGURE 10.6

The spread of infection in the network when different nodes are protected.

infected person (or to k infected people) will become infected. However, real diseases (or viral videos or information) usually do not spread that way. Some people have more robust immune systems that make them less likely to catch a disease. Some people are genetically more likely to get sick. And some people have more prolonged exposure to the sick person, increasing their risk. Furthermore, diseases spread at different rates. Some, like whooping cough, are very contagious, where almost 90% of susceptible people exposed to the sick person will catch the disease. Other diseases, like HIV, are far less contagious, with a transmission rate of less than 2% in a given incident of exposure. How can these differences be considered?

Stochastic models introduce probabilities into the models. Let p be the probability that a disease is transmitted from an infected person to a susceptible person at a given time step. The value for p will range from 0 to 1, where 0 means there is a 0% chance of transmission, and 1 means a 100% chance of transmission. If $p = 0.6$, the chance of transmission is 60%.

Consider the small network in Figure 10.7 as an initial example. Node A starts out as infected, and nodes B, C, and D are susceptible. We will assume a 1-threshold model, and let $p = 0.8$ (an 80% transmission rate). At time $t = 1$, node B can be infected by node A. Since $p = 0.8$, there is an 80% chance that node B gets infected. At time $t = 2$, node C can become infected *if* node B was infected at time $t = 1$. What is the chance that node C becomes infected? There are two possibilities. If node B was infected, there is an 80% chance it will pass the disease on to node C. If node B was not infected (which happens 20% of the time), there is no chance node C can catch it, since node B does not have the disease. Thus, there are three possible things that can happen:

- Node B is infected and passes the disease on to C.
- Node B is infected and does not pass the disease on to C.
- Node B is not infected and thus cannot pass the disease on to C.

To find out how likely it is that node C becomes infected, the probability of each option has to be considered. There is an 80% chance that node B was infected. If that happens, there is an 80% chance it infects node C. To come up with the probability that node C is infected, we multiply the chance that node B was infected by the chance of passing on the disease. Let PI be the probability of infection, and PI(B) be the probability of infection for node B. Then we can calculate PI(C) as follows:

$$\begin{aligned} \text{PI(B)} * p &= \text{PI(C)} \\ 0.8 * 0.8 &= 0.64 \end{aligned}$$

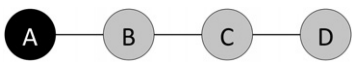


FIGURE 10.7

A small network, where node A begins as infected and nodes B, C, and D are susceptible.

The other two possible outcomes also have a probability of happening. The probability that B is sick but does not pass on the disease is:

$$\text{PI(B)} * (1 - p)$$
$$0.8 * (1 - 0.8) = 0.8 * 0.2 = 0.16$$

And the probability that B is not infected is 0.2. If we add the probability of B being infected and passing the disease (0.64), the probability of B being infected and not passing the disease (0.16) and the probability of B not being infected (0.2), they sum to 1. That means there is a 100% chance that one of these things happens, and that should make sense. These three outcomes are the only three options, so there should be a 100% that one of them occurs.

This simple network leads to the fundamental idea behind the stochastic model. The probability that a node becomes infected is given by the probability that its neighbor is infected multiplied by the probability of transmission.

Following that logic, the probability of node D becoming infected follows clearly. Node C is the only node that can infect D. The probability that C is infected is 0.64. The probability of transmission is 0.8. Thus, the chance node D is infected is:

$$\text{PI(C)} * p \quad = \text{PI(D)}$$
$$0.64 * 0.8 \quad = 0.512$$

This is the simple case, where each node has only one infected neighbor. What if there are two infected neighbors? This is shown in Figure 10.8.

Nodes A and B are infected and node C is susceptible. For now, assume a 1-threshold model and the probability that nodes A and B are infected is 1 (a 100% chance of infection). Let $p = 0.6$. What is the chance that C becomes infected? There are three scenarios where C can be infected in a 1-threshold model:

- Node A passes the disease, Node B passes the disease.
- Node A passes the disease, Node B does not pass it.
- Node A does not pass the disease, Node B passes it.

Note that the other possibility is that neither node passes the disease, but in that case, node C does not get infected.

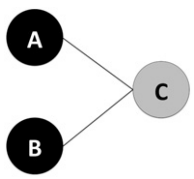


FIGURE 10.8

A network where nodes A and B are infected and node C is susceptible.

Another way to look at this is to write out all the possibilities in a tree format. The probability of each event is written in the chart. This is shown in Figure 10.9.

The probability that C is infected is the sum of the probabilities of each possible scenario listed above. If the probability of transmission (p) is 0.6, the probability that a node does not transmit the disease is $1 - 0.6 = 0.4$. These two numbers are all that are necessary to compute the probabilities above:

- Node A passes the disease, Node B passes the disease $0.6 * 0.6 = 0.36$
- Node A passes the disease, Node B does not pass it $0.6 * 0.4 = 0.24$
- Node A does not pass the disease, Node B passes it $0.4 * 0.6 = 0.24$

Note that in Figure 10.9, these probabilities are shown at the bottom under each scenario, and they are the product of the probabilities shown on each edge.

Thus, the chance that C is infected is $0.36 + 0.24 + 0.24 = 0.84$.

If this same network is considered with a 2-threshold model, then both nodes connected to C must be infected and pass on the disease in order for C to catch the disease. The only scenario where this happens is the first one listed above: when both A and B pass the disease. The probability of this happening is 0.36.

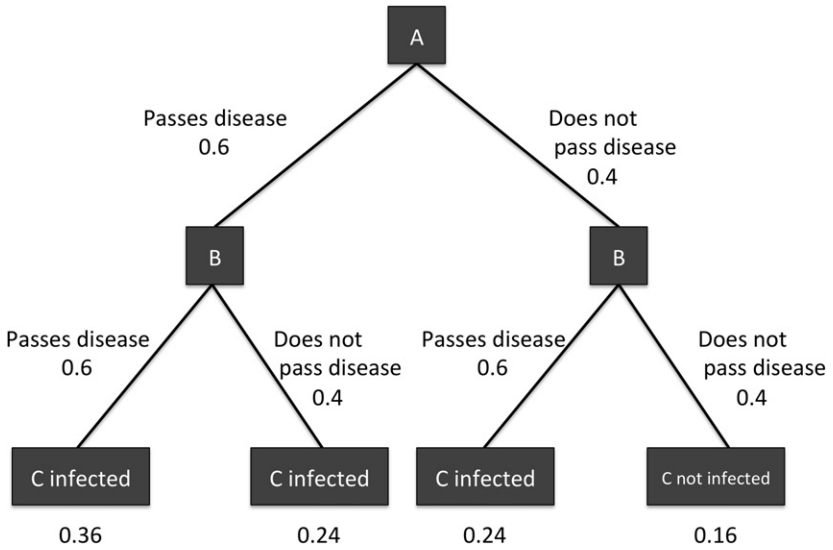


FIGURE 10.9

A tree showing the possibilities of infection from node A, and then from node B. The bottom row shows all four possibilities and the probability of each happening. Note that the probabilities on the bottom are the product of the values on the edges leading to that option.

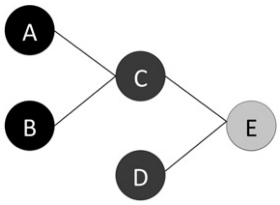


FIGURE 10.10

An extension of the network in Figure 10.8. Black nodes are definitely infected, light gray nodes are susceptible, and medium-gray nodes are infected with some probability.

None of the other scenarios will lead to infection in a 2-threshold model, so the chance that C becomes infected is simply 0.36.

In this example, there was a 100% chance that nodes A and B were infected. However, the first example in this section included nodes that were not necessarily infected. The next step is to incorporate those chances into this network. Figure 10.10 shows a network that extends the one considered so far. It includes node C, which has a 0.84 chance of being infected, and two additional nodes: D and E. Node D also has a probability of being infected. Let PI(D) = 0.7. Knowing this, what is the probability that E becomes infected?

In a 1-threshold model, there are several more scenarios that will lead to infection. First, we compute the probability for each of those scenarios. That is the probability of the scenario for C (infected or not) times the probability of the scenario for D (infected or not):

- Nodes C and D are both infected: $0.84 * 0.7 = 0.588$
- Node C is infected but node D is not: $0.84 * 0.3 = 0.252$
- Node C is not infected but node D is: $0.16 * 0.7 = 0.112$
- Neither node is infected: $0.16 * 0.3 = 0.048$

Note that node E cannot be infected in the last case, so that scenario is not carried into the calculations below. For each of the first three scenarios, the probability that the disease is transmitted must be calculated.

If C and D are indeed infected, the same three scenarios presented in the scenario with A and B infecting C apply:

- Node C passes the disease, Node D passes the disease $0.6 * 0.6 = 0.36$
- Node C passes the disease, Node D does not pass it $0.6 * 0.4 = 0.24$
- Node C does not pass the disease, Node D passes it $0.4 * 0.6 = 0.24$

So, if C and D are infected, the chance that the disease is passed to node E is 0.84. However, nodes C and D are not necessarily infected. Thus, we have to multiply 0.84 by the probability that both nodes are infected.

$$0.588 * 0.84 = 0.494$$

If node C is infected but node D is not, the chance that C passes the disease is simply p (0.6). So the total chance of the disease passing in this case is the chance of the scenario happening times the chance of infection:

$$0.252 * 0.6 = 0.151$$

The same applies to the scenario where node D is infected, but node C does not:

$$0.112 * 0.6 = 0.067$$

To get the final probability of node E becoming infected, we add the probability of it being infected when both nodes are sick, plus the probability of being infected if only C is sick, plus the probability of being infected if only D is sick:

$$0.494 + 0.151 + 0.067 = 0.712$$

The full list of probabilities of infection and transmission are shown in Figure 10.11. While this is more complicated than Figure 10.9, note that the probability shown for each scenario at the bottom is still the product of the probabilities shown on each edge leading to that scenario.

In a 2-threshold model, the case is again simpler. The only way to be infected is if both C and D are infected and if they both pass it. The chance that both nodes are infected, as computed above, is 0.588. The chance that both nodes pass the disease in this case is 0.36 (again, calculated above). Thus, the chance that node E is infected is:

$$0.588 * 0.36 = 0.212$$

These calculations can be carried forward many steps into the network, and done in networks where nodes have a higher degree. Things also become much more complicated in situations like the scenario in Figure 10.12.

In this example, nodes A and B start infected. Nodes C and D can be infected from them in time step $t = 1$, but in subsequent time steps they may be infected by one another as well. This greatly complicates the situation. In time step $t = 2$, node D could have been infected at $t = 1$, but if not, could be infected by C if C was infected at $t = 1$. At time $t = 3$, D could have been infected at $t = 1$ or 2, or it could be infected by A, B, or C at time $t = 3$ excluding the possibility that C was infected by D at time $t = 2$.

In a large network, these problems can become overwhelming. Generally, in networks with any complexity, exact probabilities are not computed as we have done here, but rather simulation is used to show how the disease would spread through the network at each time step.

Simulation is a complex topic beyond the scope of this chapter, but a brief discussion will highlight many of the issues presented so far. Simulations use rules, mathematical guidelines, and probabilities to assess what would happen under real conditions. A simulation of disease spreading through a social network would include a compartmental model (SI, SIR, etc.), a k-threshold value, and

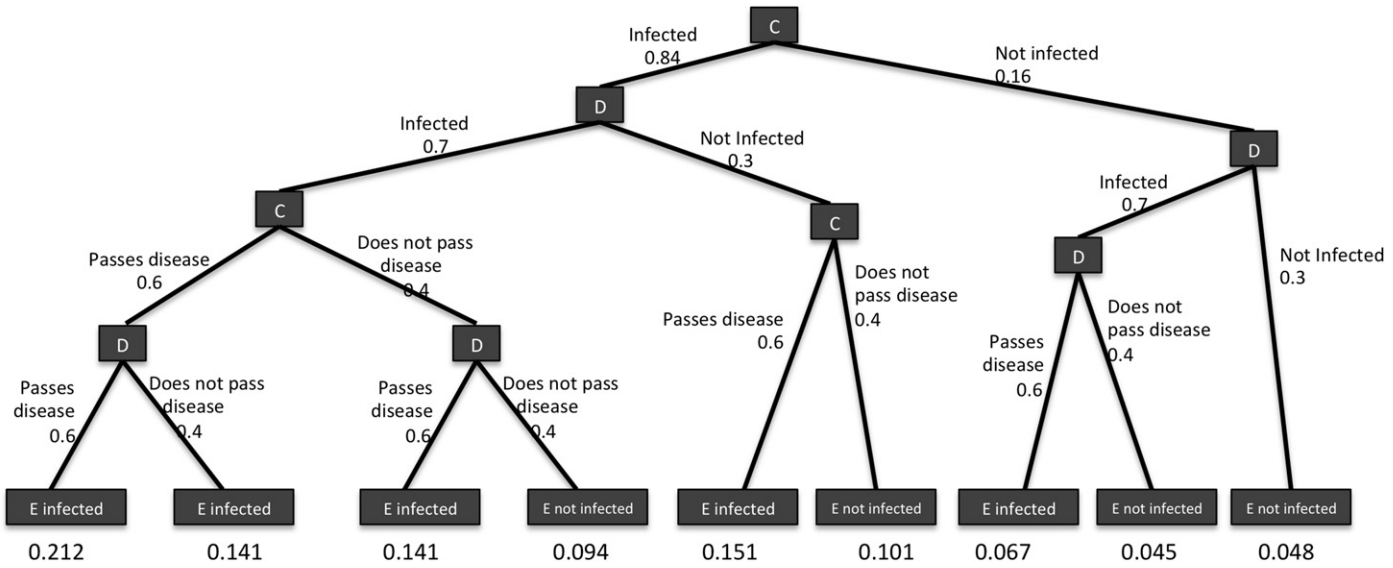


FIGURE 10.11

The full set of scenarios for nodes C and D being infected and passing the disease to node E.

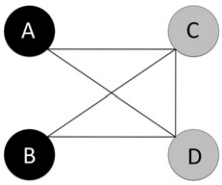


FIGURE 10.12

A network where nodes C and D can be infected by nodes A and B, or by one another.

probabilities of disease transmission and infection. Scenarios may become more complex than those we discussed; for example, some nodes may be more at risk than others, and these differences could be modeled in the simulation.

Applications of epidemic models to social media

Epidemic models, like those presented above, were originally designed to study the spread of disease through populations. However, they are now used to model the spread of many other things in social media. This section presents one example.

Viral marketing is a topic of great interest to companies, especially online retailers, because it spreads information about products at no cost to the seller. This is good if the product reviews and feedback are positive, but it can backfire if negative information spreads virally.

Researchers (Leskovec et al., 2007) investigated "viral" recommendations that users send to one another and their impact on purchasing behavior. Nodes in the network were users of an e-commerce system, and there was an edge between users if one recommended an item to the other. The researchers used an SIR model, where susceptible individuals have not purchased an item, an "infected" person buys an item based on a recommendation, and after purchase, a person is "recovered" and not susceptible to buying the item again.

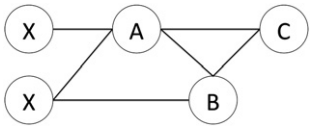
Using real data from the system, the researchers looked at a threshold model (more complex than the basic k-threshold model presented above) to understand how recommendations spread in a network. They found that, overall, recommendations did not spread much in a viral way, but for some products the recommendations did have more of a reach into the network.

Exercises

1. Work the Firefighter Problem on the example network where you are allowed to place five firefighters each round in a 1-threshold model. Assume

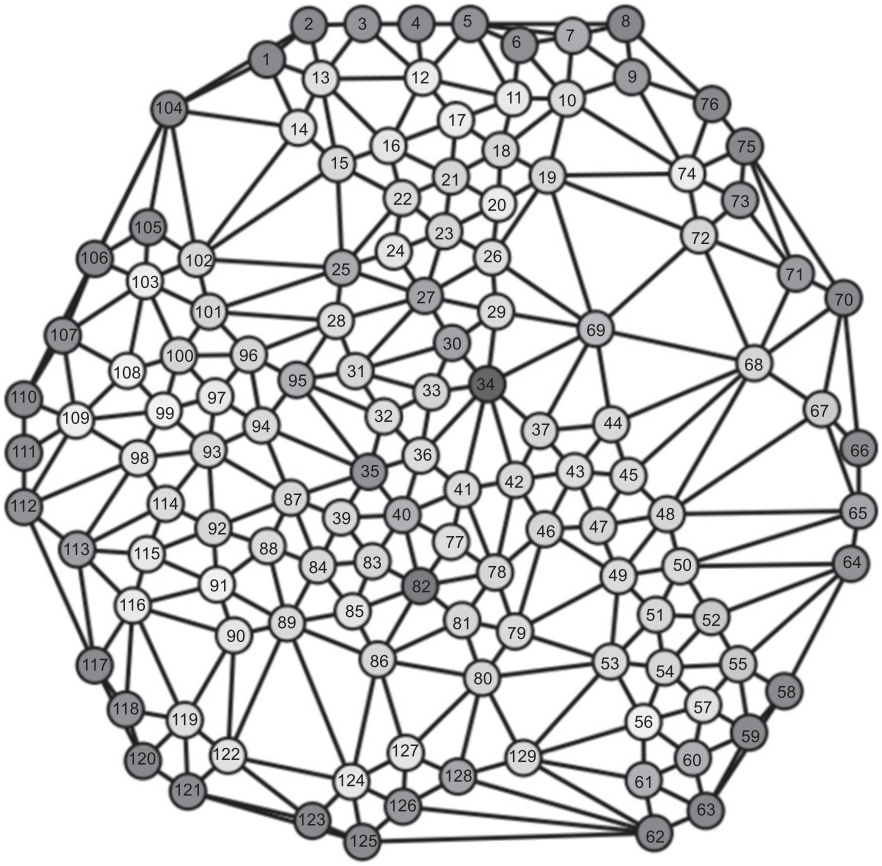
all nodes whose number ends in a "5" begin infected. Try two different solutions and compare the number of uninfected nodes left at the end of each process.

2. Again, assume all nodes whose number ends in "5" begin infected. Using a 1-threshold SIR model, trace the spread of a disease in the network. A node will remain infected for one time step and then become recovered. A node can only be infected by a neighbor in the infected "I" state.
   - **a.** How many time steps does it take before the disease stops spreading?
   - **b.** What percentage of nodes end up in the R state? In the S state?
   - **c.** How does this change using a 2-threshold model?
3. Consider this graph. Nodes marked with an "x" are definitely infected at time $t = 0$.

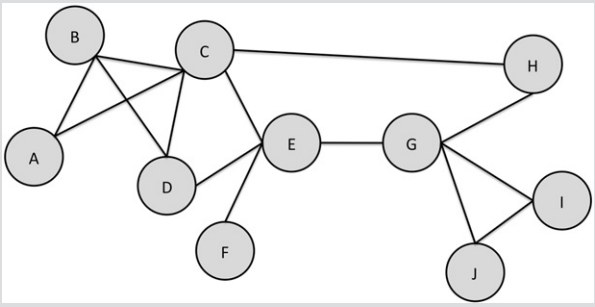


   - **a.** For each node A, B, and C give the probability that it is infected if p = 0.7 and k = 2 at time $t = 1$ and $t = 2$.
   - **b.** For each node A, B, and C give the probability that it is infected if p = 0.7 and k = 1 at time $t = 1$ and $t = 2$.
   - **c.** For each node A, B, and C give the probability that it is infected if p = 0.5 and k = 2 at time $t = 1$ and $t = 2$.
4. You are in charge of creating a viral marketing campaign for a company. Imagine you have access to the entire Twitter social network. Which Twitter users would you target with information about your campaign with the goal of having your message spread as widely as possible? Give specific social network statistics you would look for in target users and, based on the material covered in this chapter, explain your choices.
5. In the large graph below, imagine you have an SI, 1-threshold model.
   - **a.** Which node would you infect to reach the greatest number of people? How many people become infected if you choose that node?

b. Imagine node 34 begins as infected. The disease will spread and then you can inoculate two nodes to prevent them from becoming infected. These must be nodes that are not already infected. After your inoculations, the disease will spread again, and then you can inoculate two additional nodes. Repeat this process until the disease can no longer spread. What percentage of nodes become infected using your inoculation strategy? Compare your results to those of other students.



INFECTION PROPAGATION



In the graph above, consider how infection propagates. For example, imagine a 2-threshold SI model where nodes I and J begin as infected. At time step 1, node G becomes infected. Nodes I and J remain infected, and then the infection stops spreading. Three of the 10 nodes are infected at that point, or 30% of the network.

Depending on the threshold and model, the spread of the disease will change dramatically. In some cases the disease will spread to only a few nodes. In some cases all nodes will be infected and, depending on the model, will either remain infected or recover. And in other cases, the disease will cycle through the network, continuously re-infecting nodes that had previously recovered.

Assume nodes C and G start infected. For each of the model-threshold combinations below, determine if the disease continues to circulate or if it stops. If it stops, calculate the percentage of nodes who had been infected (even if they eventually recovered).

Fill in the Percentage of Nodes that become Infected for Each Scenario

	1-threshold	**2-threshold**
SI		
SIR		
SIS		
SIRS		

CHAPTER

Community-Maintained Resources

11

In the offline world, many resources are available to the public—parks, public libraries, community gardens, and museums are just a few examples. The people who go to these places and the community that rises up around their interactions often define what kind of place it is to visit. The users of the facilities are a very important part of the culture.

Many such places are maintained by municipalities (e.g., cities, park districts, counties, etc.) or private companies. They set the policies, take care of the facilities, and regulate their use. However, in some cases, as with community parks and gardens, the community members themselves take on the administrative roles. They use the facilities but also help manage them at every level.

The same is true online. In social media, community members are responsible for generating most or all of the content. Whether it is message boards, social networks, blogs, or social bookmarking sites, users are making posts, commenting, and generally filling the sites with information.

But on some sites, users play a role in managing the websites that host their activities. They may monitor the quality of the content, set policies, and perform administrative functions. *Community-maintained resources* are websites where community members—often all volunteers—are responsible for the full management and maintenance of the site, and where the content provides some service or resource to the community members. This chapter will discuss the technologies and user motivations that support the existence of these sites.

Supporting technologies for community-maintained resources

Community-maintained resources are interesting for what they offer users and the role that users play in them. Thus, this chapter will focus on sites and services that offer a resource to users and where the users are content consumers, content producers, managers, auditors, policy makers, and administrators.

A community-maintained resource can exist in any website, mobile application, or online service. Participants can generate everything from the content to the interface features to the underlying code. However, most resources are built on top of collaborative software, sometimes called groupware. The software itself

does not necessarily qualify something as a community-maintained resource, but some applications are specifically designed to support them.

Wikis

Wikis are a prime example of groupware that supports the creation of community-maintained resources. Wikis are websites where users can create, modify, and delete content. They use a simple markup language for formatting, so users do not need to know HTML (the language used for authoring web pages from scratch).

The first wiki was designed by Ward Cunningham in 1994 and designed for quick web editing. He got the name from a quick airport shuttle service at Honolulu International Airport called the Wiki Wiki Shuttle and used it to name his software WikiWikiWeb.

There are now many types of wiki software. These applications are installed onto a web server and they handle the storage of files, revision history, user information, and other back-end data. Wikis are created using this software.

Community members contribute to wikis in every way. They generate the content, revise the pages, have and moderate discussions about the pages, and establish the scope of the wiki.

At the time of writing, the most popular wiki is Wikipedia, the community-authored encyclopedia, which is one of the 10 most visited websites in the world. Other popular examples include WikiHow, focused on how-to instructions; WikiAnswers, a question answering website; Wiktionary, an online dictionary with meanings and pronunciations of words in every language; and domain-specific wikis, such as Lostpedia or HRWiki, which attempt to collect comprehensive information on a specific topic (for these examples, the TV show *Lost* and the web comic Homestar Runner, respectively).

Message boards

Some of the earliest ways people created communities and resources online was through message boards. One of the first of these was Usenet, which predates the web. First established in 1980, Usenet was a collection of message boards (called newsgroups) where users could post questions and articles and have discussions. Many Usenet boards became resources for people to find reference materials and get answers to questions from experts on a given topic.

Message boards evolved into the online forums and discussion boards that are available by the thousands on the web today.

Many message boards are excellent resources, with answers to questions, pointers to resources on a given topic, news, solutions to problems, or just as places for people to have discussions with others in an online community.

Message boards or forums need not be community-maintained nor resources. Some are maintained and run by companies or other organizations who dictate

the policies, account restrictions, and activities, independent of the community. Furthermore, they may not be resources at all. Some boards are casual places for chat with no sense of community, few returning members, and no artifacts left behind that may be useful to others.

Repositories

A common way communities build shared resources is through building and curating collections of useful items.

One large use of repositories is in the open-source software community. Communities of programmers build collections of code. For example, people who use the Perl programming language contribute "modules," which are packages of code to perform specific functions, like accessing web pages, doing statistical analysis, or parsing text. Their community website, CPAN (cpan.org), is a repository for storing, managing, discussing, and providing access to the modules.

In a totally different domain, Ravelry (ravelry.com) is a community for knitters and crocheters. Among the social features, they maintain a repository of patterns contributed to and organized by users.

As with wikis, many types of repository software can be installed on servers and used to support a specific repository, or websites may choose to create their own repository code to support their specific needs.

User motivations

Within community-maintained resources, people play different roles. Some—often the majority—are casual users who use the resource without contributing anything back. What motivates users and how the design of the community contributes to its success are all important factors for understanding successful community-maintained resources.

Karau and Williams (2001) have developed a model to explain individual motivation in groups called the Collective Effort Model (CEM). Their model asserts that individuals are willing to contribute to a task that benefits a group only if they believe their work will lead to outcomes that they personally value. Those outcomes may be activities that help them earn money, that keep the resources current, accurate, and useful for others in the hope that others will do the same in return, or that are valuable for serving others who need help. Furthermore, the CEM puts forth if a person cannot see meaning or value in the outcome of an activity, or if it is not clear that their efforts will lead to valuable outcomes, they are unlikely to participate.

The remainder of this section presents two case studies on user motivation. The insights from the CEM are present in all of the examples discussed.

User Motivation—case study: Wikipedia

Background

Wikipedia (http://wikipedia.org) is a user-created, -edited, and -maintained encyclopedia built using the Mediawiki software. As of mid-2012, it is available in 285 languages, and the English version is the most popular. The English version contains roughly 4 million articles, all created and edited by volunteers. We will focus on the English version of Wikipedia in this section.Although there are disputes about the quality of the articles on Wikipedia, it aspires to be a neutral, authoritative, verified, well-balanced, and complete source of information.

Anyone can edit or create an article on Wikipedia. Much of the maintenance of the site happens through small edits from readers. However, the site has an explicit hierarchy of user roles that grants additional privileges to people who have moved up within the community. Editors, both registered and anonymous, can make changes from correcting spelling or punctuation to deleting, rewriting, and adding large sections to articles. At time of writing, there were roughly 85,000 active contributors (those who had edited within the last month). As editors become more active and trusted within the community, they can be promoted to administrator positions. There are nearly 1,500 administrators, and they have the power to block or ban users and give additional permissions to editors. They are nominated and approved by the highest ranking and smallest group of volunteers, the Bureaucrats. There are only 34 of these users. They have only two main functions: the appointment of administrators and the selection of other bureaucrats.

The higher-level statuses are reserved for a relatively small number of editors. The vast majority of all the content is created and edited by nonadministrators.

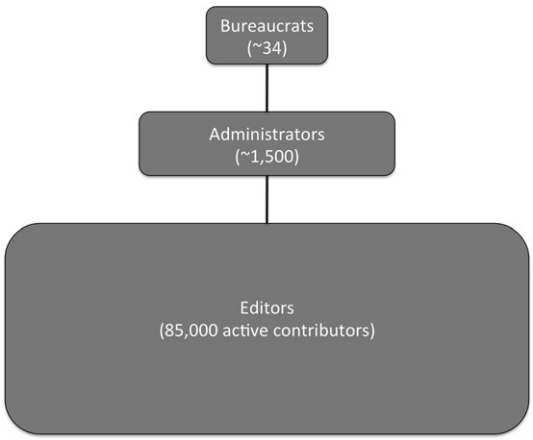


FIGURE 11.1

The structure of the editing community within Wikipedia.

This keeps the site running. With over 3 million edits per month made on the site, it can only keep functioning through the efforts of the average user. Thus, while there are explicit user roles in Wikipedia, a variety of roles and participation levels will also be present among the large group of editors.

Editor motivation

What motivates people to participate in editing and writing for Wikipedia? They are not paid for their work, and the wiki-based nature of the website means they do not even receive public credit for their efforts.

The answer is that there is no single motivation for people to edit Wikipedia. Researchers have conducted a number of studies on the topic, and they find that authors provide a wide range of reasons for contributing. There are some common themes that emerge.

Forte and Bruckman (2005) found that receiving credit for one's work and building reputation in the community are important factors for very active editors. While Wikipedia does not list contributors to articles, and the community standards prevent people from claiming authorship, the wiki technology reveals who has created and edited articles. Contributors can build up reputations as active editors or discussion participants, particularly if they are working on a number of articles related to a given topic. This builds their reputation among other editors and earns them credit, even if that is not made explicit to readers. This type of credit and reputation can also be used to earn a promotion within the hierarchy of Wikipedia to become an administrator, and thus it has value for people who want to advance within the volunteer organization.

Nov (2007) conducted a survey of Wikipedia contributors and asked them questions about their motivation. These motivations were tied to previously identified volunteering motivations that included the following:

- Values—wanting to help others
- Social—allowing people to engage with others and receive reputational credit for participating in a good activity
- Understanding—learning new things through volunteering
- Career—learning skills that may help with finding, keeping, or advancing in a job
- Protective—reducing guilt about one's privilege by sharing with others
- Enhancement—serving the community (similar to the protective motivation, but without the guilt component)
- Fun—finding the activity enjoyable
- Ideology—believing that information should be freely available

Users were motivated by each of these factors, but by far the most important were fun and ideology.

Yang and Li (2010) also found that fun was an initial motivator for Wikipedia editors, but that over time, editors continue to contribute because feelings of personal achievement come along with editing.

Finally, Glott et al. (2010) surveyed people about why they contribute to the site. Answers touched on all the major ideas listed above, but users mentioned two major factors. Nearly 73% of respondents said they like the idea of sharing and wanted to contribute, and 69% said they saw an error that they wanted to fix.

Thus, while editors all have their own reasons for contributing, the major motivations for editing that emerge from all of these studies are building reputation within the Wikipedia community, participating in a fun activity, improving the resource, and enjoying the feeling of personal achievement. These lessons can carry over to help understand the motivation for participating in many different community-maintained resources.

Site maintenance—case study: Geocaching

Background

Geocaching (http://geocaching.com) is a treasure hunting game. Participants (*cachers*) hide waterproof containers (*caches*) outside and post the GPS coordinates online. Other cachers use GPS-enabled devices to search for the caches and, when they find them, they log their achievement on the website. Geocaching is a community-maintained resource because all of the caches are created, hidden, and maintained by participants.

Small caches may be the size of a magnetic key holder, mint tin, or film canister; medium sizes are often Tupperware-style containers; and larger caches are in boxes like ammunition cases. A cache typically contains a logbook that people can sign when they find it, and small trinkets for cachers to take and exchange. Although the coordinates of a cache are given, the cache itself is usually well hidden in that location. Figure 11.2 shows an example of a cache in its hiding place and its contents.

The first geocache was placed in 2000. As of mid-2012, there are over 1.7 million caches hidden around the world (plus one on the International Space Station), and over 5 million geocachers.

Information about a cache is given on the cache's page located on the geocaching website. That same data is usually also available on GPS-enabled devices and via smart phone apps that support geocaching. Figure 11.2 shows the data for an example cache. This includes the coordinates, a map, information about the difficulty of finding the cache (in terms of how well hidden it is and how difficult it is to reach), a description provided by the creator, hints, and logs from cachers who have searched for it.

Reasons for participating in geocaching vary. A study by O'Hara (2008) found a number of major motivations.

- Walking—Going out to find a geocache provided motivation for people to go walking that they would not have had otherwise. Parents in particular indicated that it was a way to make their children interested in going outside and engaging in some physical activity that they would not want to do without the caching motivation.

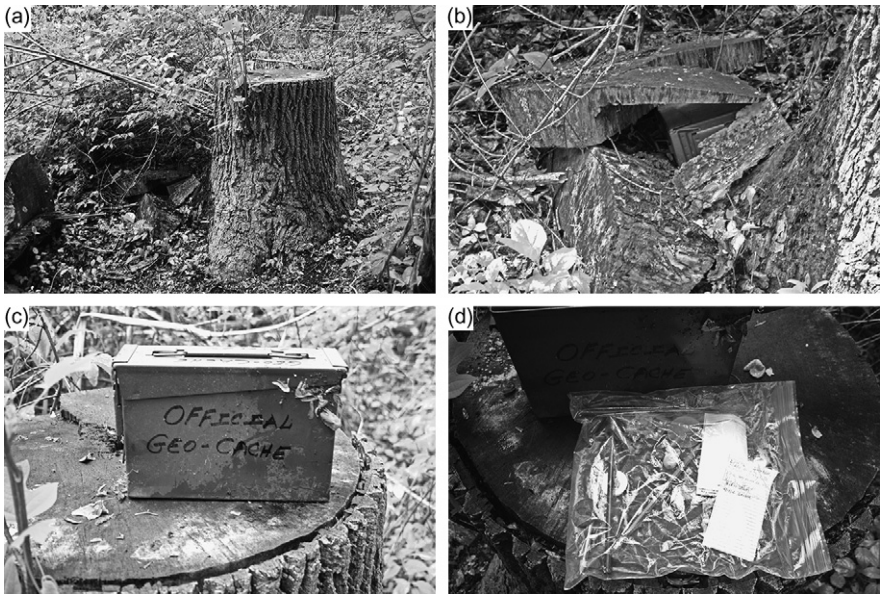


FIGURE 11.2

An example of a hidden geocache. The edge is peaking out in (a). A zoomed-in view to the left side of the stump is shown in (b), with the cache more clearly visible. The cache, now removed from its hiding place, is shown in (c). Contents, including a sealed plastic back with a log sheet, pencils, and trinkets, are shown in (d).

- Exploration—Caching is a way to explore cities, parks, open spaces, and new neighborhoods. Geocachers explained that searching for caches gave them reasons to learn and discover new things about their communities, and to go to places they might not otherwise explore.
- Logging caches—When a geocacher finds a cache, she logs it in the cache's logbook and on the geocaching website. This motivates people in several ways. First, there is an aspect of collecting to it, where people expand the list of caches they have found. Related to this is the notion of statistics. In addition to a list of caches found, a user's profile shows the number of caches they have found and hidden. Increasing that number and improving their geocaching profile are motivations for finding more caches.
- Competition—Within the community, being the first person to find a cache is a small honor. Cachers will often go out as soon as a new cache location is posted to try to be the first to find it. This competition to get there first drives some participants.
- Challenge—Basically all geocaches have some challenge to them, whether it is finding them in their hiding places at the given GPS location, or getting to the location itself. Succeeding at that challenge is both an individual and social motivation for many users.

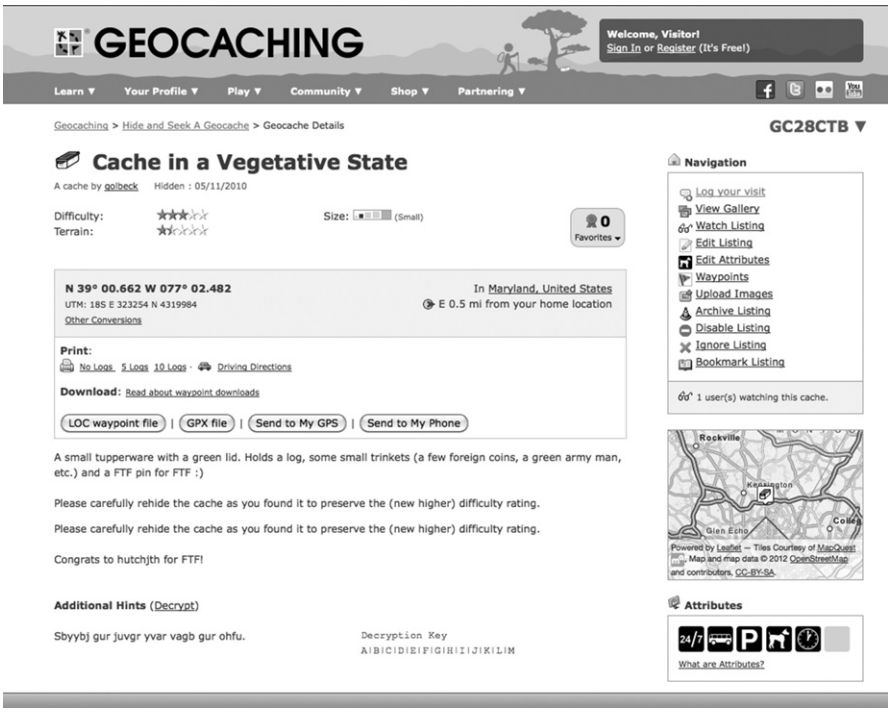


FIGURE 11.3

A geocaching page for an example cache, taken from geocaching.com.

Maintenance

The most important issue for the success of geocaching is that the caches are maintained and that new caches are made available. The caches are not created or maintained by a company or central organization; regular geocaching participants create and hide caches and are responsible for their maintenance. Once placed, the work is not done; caches need to be taken care of. It is not uncommon for caches to be damaged in weather events, to be moved or destroyed by nongeocachers (known as *muggles* within the community), or to need basic maintenance like new logbooks. Some active users will have hidden dozens of caches, and keeping them in good repair puts a burden on cache creators.

How does the structure of the community encourage maintenance of this resource? Some users volunteer as administrators, who can activate or deactivate caches and contact their owners if the cache has gone for some time in need of maintenance. However, the main source of information to help with cache maintenance—both to cache owners and administrators—is the people who search for them.

When a geocacher finds a cache, she posts a log indicating her find on the geocaching website. This indicates to the cache owner that the cache is still there

and in good condition. Geocachers may also fail to find a cache and log their non-find as well. While this does not necessarily mean a cache has disappeared, several consecutive logs of this type are a sign to cache owners that they may need to check on the cache. Searchers can also post logs indicating that a cache needs maintenance. On occasion, other searchers will do this maintenance when they visit the cache. For example, if a logbook is full, a fellow cacher may put a new one in. That replacement would also be noted in the website log.

Thus, through their natural behavior, cache seekers provide information that relieves much of the burden on cache creators. A cache owner hides the cache, participants find it and report back on its status, and the owner has to visit the cache only when seekers identify a problem.

In community-maintained resources, this cycle of natural user behavior supporting the maintenance of the site is often an important factor in the site's success. As Neustaedter et al. (2010) note, maintenance in geocaching is often a side effect of other participation in the community, that is, sharing experiences of finds. However, the nature of maintenance through small actions is one that carries across many community-maintained resources, like Wikipedia and others.

Exercises

1. Make an edit on Wikipedia.
2. For the Wikipedia page you edited in exercise 1, look at the edit history of the page. How significant is your contribution? How frequently is the page edited?
3. Look at the edit history of the Wikipedia page on World War II and compare it to the edit history for the Chremonidean War. Characterize the differences. How do you think this impacts the quality of the article?
4. Download the free Geocaching app to your mobile device and find a geocache near you.
5. For the geocache you found and two others, look at the history of logs. How many "found it," "didn't find it," and "maintenance needed" logs are there for each? How have the cache owners or community responded to requests for maintenance?
6. List three community-maintained resources that you use.
   a. What is your level of participation in each?
   b. Select the one that you contribute to least. What would make you participate more? If you are not motivated to participate more, why not?
   c. Select the one you contribute to most. What is your opinion of people who only use the resource but do not contribute anything back to it?
   d. Recall the Collective Effort Model. What valuable outcome do you expect to contribute to through your efforts in the community-maintained resource?

7. Find five friends who have contributed to a wiki. Interview them and summarize their answers to the following questions:
   a. How often do you contribute to the wiki?
   b. What are the main kinds of contributions you make?
   c. Why do you contribute to the wiki?
   d. What do you get out of contributing?
8. A frequent event on Wikipedia is that people delete or vandalize pages.
   a. It takes effort to vandalize the page. What do you think is the motivation of these people?
   b. Why do you think the community quickly responds to fix these problems? What is their motivation?

CHAPTER

Location-Based Social Interaction

12

Identifying location in social interactions is emerging as one of the new trends in social media. Location data is used to create new connections (both on- and offline) and to spread information about the locations themselves. It may be explicitly shared by users or passively collected by sites that use it to associate social behavior with geographic patterns. This chapter discusses the role of location information in social interaction.

Location technology

There are many different ways that location information is accessible from users. They may explicitly state their location, or the location may be automatically detected from their computer's IP address, pulled from a mobile device's GPS device or from cell phone tower triangulation, and even from leveraging wifi hotspots.

User-posted location data

In some social media websites, users list their default location. For example, in Twitter and Facebook, users can add a hometown to their profile. Figure 12.1 shows the Twitter profile page for the *Washington Post*, and its provided location (Washington, D.C.) is boxed in black.

Estimating location data via IP address

Even when users do not list a location, it is still possible to determine a nearby location from the user's IP address. Every computer on the Internet has a unique number that identifies it; this is called an IP address. An IP address has the form of four numbers between 0 and 255, separated by dots. For example, 192.0.43.10 is a traditional IP address.

Newer IP addresses use a new system called IPv6. IPv6 has more possible numbers, allowing more devices on to the Internet. These addresses are longer and look slightly different. For example, 2001:500:88:200::10.

IP addresses are given in blocks to Internet service providers (ISPs). ISPs may be cable companies that provide home access to Internet users, businesses whose

FIGURE 12.1

The *Washington Post* Twitter profile page with its location indicated in the black box toward the top.

employees all connect to the Internet at work, or universities that connect labs, student laptops, and faculty members.

Whenever a person accesses a site online, her IP address is sent to the site. Many databases will provide a location for any IP address. The databases are built from many sources of information (including regional Internet registries that show which country the IP address is assigned to, user-submitted data, data provided by ISPs, and other sources). While they are not always accurate down to the specific city where a computer is located, they are often close. This data can be used to assign a rough location to information that is posted online.

GPS location data

Finally, GPS-enabled devices (particularly mobile phones) can automatically associate very specific location information with data that users post online. GPS is the Global Positioning System. There are 30 satellites orbiting the earth, and, when a GPS-enabled device is in view of at least four satellites, the device's

position can be pinpointed to within a few feet. GPS-enabled mobile devices, cell tower triangulation, and other technologies can use the location when a user posts photos, status updates, and other content.

Mobile location sharing

Because of the ease with which accurate location can be associated with online activities on mobile devices, explicitly integrating social interaction and location has become a common and growing area of activity. Some of these location-based social activities are game-like—although people most often use location data to notify friends of their whereabouts, find friends near them, and track where they have been.

Created in 2000, one of the earliest mobile location-based applications was called Dodgeball. Users would text their location to the service, and it would notify them if any of their friends were nearby. (In this early system, GPS was not used for location identification; the system relied on users reporting their location in their text messages.) Dodgeball was purchased by Google in 2005 and discontinued in 2009. Its founders left Google after the purchase and went on to found a similar service called FourSquare in 2009.

FourSquare has over 20 million users. With the user's GPS location, it allows them to pick from a list of nearby venues (or to add a new venue). Users receive points for "checking in" to a location, can connect with nearby friends, or receive information from the places they have checked in.

FourSquare also offers "badges" that users acquire for completing "challenges" (such as checking in at five different pizza places, checking in ten times at the same place, or checking in after midnight several days in a row). Users who check in the most frequently at a certain location can also claim "mayorship" status for that place.

FourSquare has partnered with businesses to allow them to offer coupons and special offers to users who have checked in. Figure 12.2 shows a FourSquare check-in screen and place information page. Research (Lindqvist et al., 2011) has shown that the gaming aspect of FourSquare was important to some users, as was the ability to see where their friends checked in.

More traditional social networking websites (like Facebook and Twitter) have created mobile applications that are able to take advantage of the available location identification features. Facebook lets users check in at a place, just as they would on FourSquare. Both Facebook and Twitter let users attach a location to any post they make. This can be latitude and longitude coordinates, the name of a place, or a city. Research has shown that Facebook users are not using the check-in feature frequently—but when they do, it is primarily to identify that they are at a place they expect their friends might be (Strange, 2011).

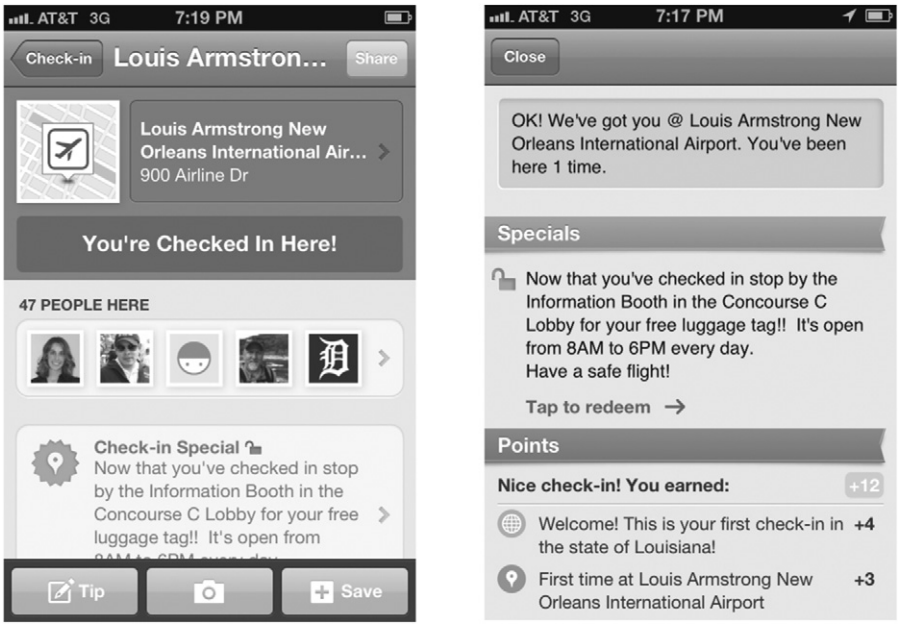


FIGURE 12.2

From FourSquare: The information page for Louis Armstrong New Orleans International Airport (left) and a check-in page (right). Notice on the left that the page includes a list of people checked in at the location, and on the right is a special offer and a section showing the points the user has earned for this check-in.

Location-based social media analysis

When users share their location voluntarily (via GPS or explicit listing) or involuntarily (estimated from IP addresses or cell-phone tower triangulation), combining location with social data opens up opportunities for many new types of analysis.

Location information can provide insights into user behavior, such as the discovery of patterns of social behavior, or how information spreads in social media. Basic analysis may include an evaluation of behavior, grouped by location. For example, Figure 12.3 shows the percentage of each U.S. state's population using Facebook.

Location data also facilitates tracking propagation across geographic spaces. Researchers at Twitter studied the flow of information out of Japan after the 2011 earthquake. The visualization in Figure 12.4 shows the flow of tweets out of Japan to other countries, and how they were subsequently re-tweeted to further locations.

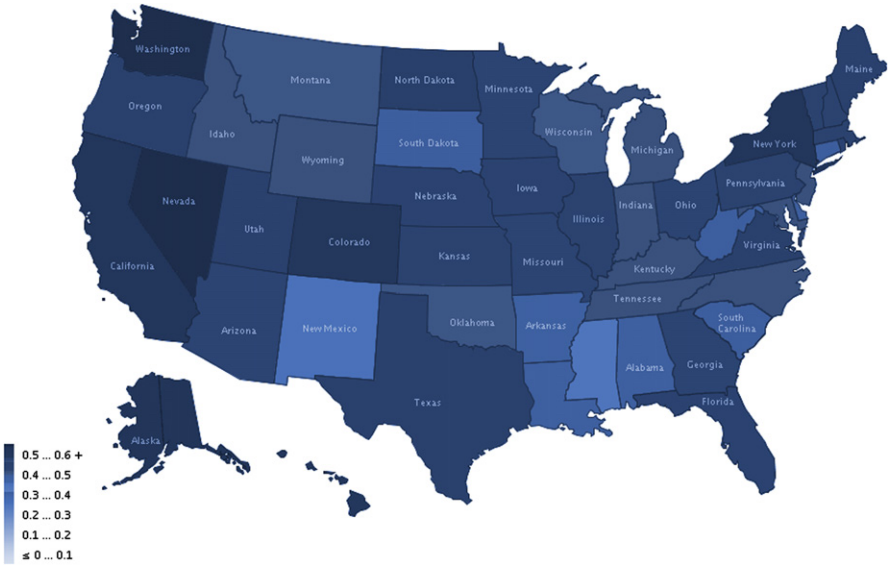


FIGURE 12.3

Percentage of Facebook users per state.

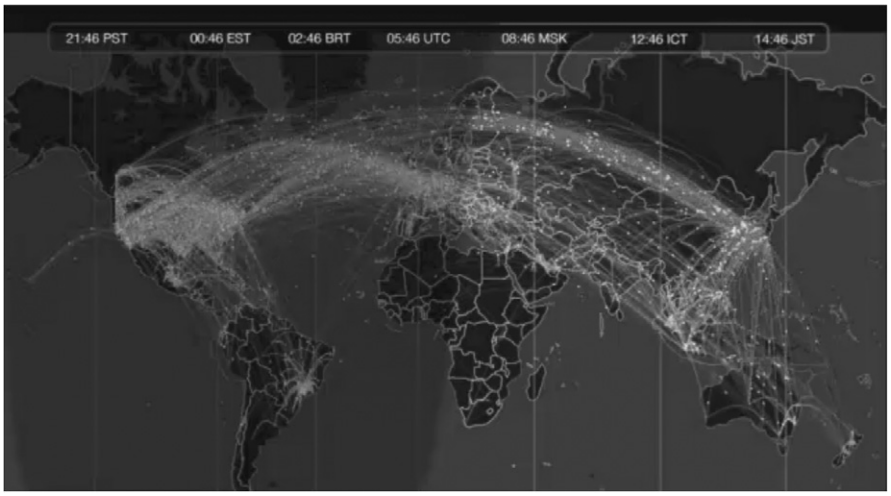


FIGURE 12.4

The flow of tweets out of Japan immediately following the 2011 earthquake, and then the re-tweets of those messages. Each arc represents a tweet flowing from one location to another.

Location-based analysis of offline events

The flu

Figure 12.4 illustrates user behavior in social media, but behavior can be used for offline phenomena, too. Monitoring the flu is one of the more commonly tracked events. Outside the purely social space, search trends from Google, paired with location information, have been shown to predict flu outbreaks before clinical organizations detect them (Carneiro and Mylonakis, 2009; Valdivia et al., 2010). When users search for "flu" or related terms, their search results paired with their location may indicate where a flu outbreak is happening.

Social media has also proven effective at showing flu trends. Lampos and Cristianini (2012) correlated reports of flu and related symptoms on Twitter with flu rates from the UK's Health Protection Agency (HPA). The researchers collected a large body of tweets each day, and analyzed them for flu-related words (such as fever, lungs, unwell, and headache).

They combined measures of these words' frequency and importance to indicate a possible flu outbreak. This yielded a "flu score" for each day. Figure 12.5 shows the relationship between flu rates as reported by the HPA and those inferred from Twitter for a specific region of the UK. There is a very strong relationship, showing that Twitter content accurately reports flu outbreaks.

Knowing the user's location is critical for this kind of analysis. Without it, flu rates around the world can be monitored, but information about outbreaks in specific locations would be lost.

Fires

Researchers have also looked at using Twitter to understand crisis events. Longueville et al. (2009) studied the relationship between Twitter posts and a large forest fire in France. They found that generally tweets accurately reflected

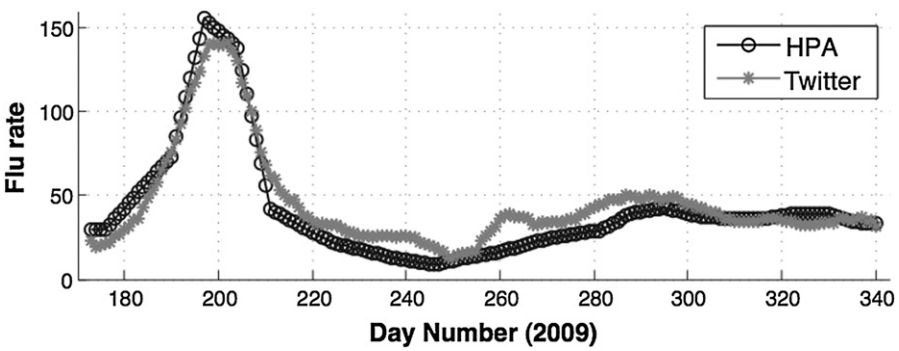


FIGURE 12.5

The relationship between HPA observed flu rates and those inferred from Twitter.

Figure and results from Lampos and Cristianini (2011).

the events happening in real time and that the location of the tweets could provide good data about emerging events (e.g., "I can see flames from my house," paired with the poster's location).

Another example of how social media is used to monitor wildfires can be found in the Colorado Wildfire Viewer.[1] It integrates Twitter posts, images from Flickr, and YouTube videos (all linked with location), along with RSS feeds (a format for syndicating web content, frequently expanded as Really Simple Syndication) and other data to produce an interactive map of the activity. Figure 12.6 shows a screenshot from the summer 2012 Hyde Park wildfire. Note the icons for Flickr and YouTube. When clicked, they reveal previews of the corresponding image or video. Twitter posts are aggregated by location. The city of Fort Collins on the eastern edge of the fire has a circle with a "33" in it, indicating that there are 33 new tweets about the fire posted by users who have listed Fort Collins as their home location. There are also two tweets posted from the Western edge of the fire, with the shared location of Larimer County, Colorado.

Tools like this allow users to keep up to date with new user-generated content about unfolding crisis situations. As with the flu outbreak data, having accurate

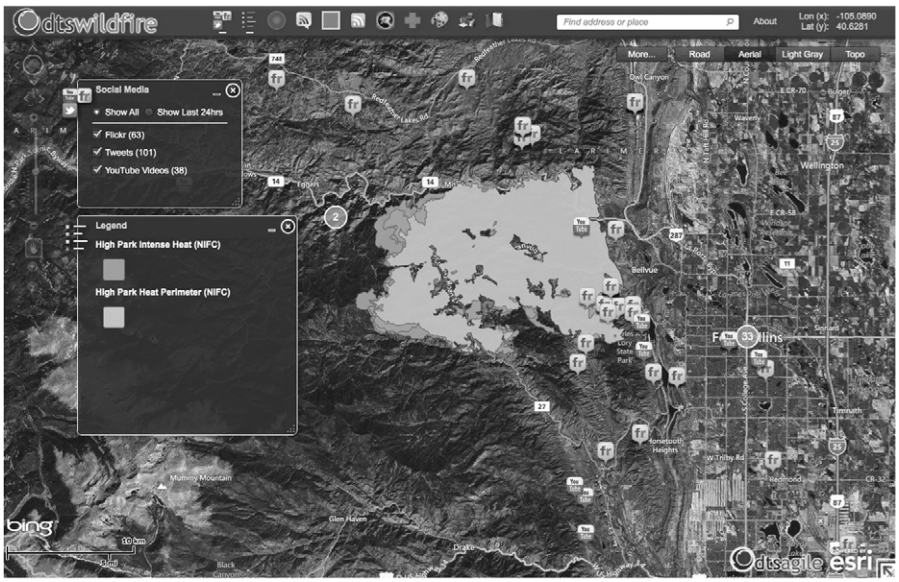


FIGURE 12.6

A mashup of Twitter, Flickr, and YouTube media combined with location surrounding the summer 2012 Hyde Park wildfire in Colorado.

[1] http://co.dtswildfire.com

location information is important. News outlets and users from areas far away from an unfolding event may post about it. Thus, having location information can help identify first-hand accounts and updates.

Crowdsourced crisis information

Ushahidi is an open-source software tool designed to support people in sharing information about ongoing crises. It began in the aftermath of the 2007 Kenyan presidential election. Both the incumbent and challenging parties were alleged by international observers to have engaged in manipulation of the election, and two months of violence followed throughout the country. Ushahidi created a website where people could report incidences of violence via email or text message. Volunteers then place reports on a map.

The information on Ushahidi was reported to be more timely and accurate than press reports or government data. Since that initial application, the software has been used in a number of applications. Of particular note is its application to natural disasters. It was used notably in response to the Hatian earthquake in 2010 as well as for earthquakes in Chile and New Zealand. People could tweet, text, or send Skype messages about people trapped or places where assistance was required. Volunteers would plot these requests, repost them on social media, and then post responses when they were received.

Marketing

Marketers are also very interested in taking advantage of mobile, location-based social media. FourSquare already has some of these features. It will show users special offers at places they have checked i, and will also indicate when businesses nearby have offers available with FourSquare. Some social coupon companies have developed or purchased companies with location-based capabilities. QR codes, square bar codes that can be scanned with cameras in mobile devices, are used to help mobile users easily bookmark a web address. Marketers use different codes in different locations to track where website visitors have seen their advertisements.

While we don't yet have an environment where our mobile devices make offers and send ads to us every time we walk past a store that wants to catch our attention, the technology to support that is here. Fortunately, privacy controls around location-based social services are fairly good. Most applications make location sharing an opt-in service, and real-time information about a person's location is usually shared with a limited number of people. It remains to be seen how well privacy will be handled as location-based services become more extensive and more integrated into social interaction.

Privacy and location-based social media

While users tend to reveal a lot of information about themselves in social media, few things are as sensitive as location. It could be a home address or a current

location that reveals the user is away from home. People with bad intentions could use location information for stalking or to break into the user's home when the user is away. Because this information is so sensitive, privacy controls are extremely important, as is awareness of who has access to location information and how it might be used.

In the early days of FourSquare, users often set up their accounts to post information about their check-ins to their Twitter accounts. This would let everyone who followed them (and often everyone on the Internet) know when they had checked in to a new location. While it may seem fun to share the great restaurant or vacation spot we're enjoying, there are risks that come along with it. A website calling itself "Please Rob Me"[2] set out to highlight those risks for users.

Please Rob Me would automatically search Twitter for any posts automatically created by FourSquare. They would then post a list of "opportunities" for each person who had just left home. Since many people use their real name and hometowns in their Twitter profile, it is a small step to go from one of these listings to a white pages website to get an address.

Please Rob Me did not have malicious intent; its intention was to educate people about the risks involved in location sharing. The site has since stopped posting these lists of people who had checked in away from home, but Figure 12.7 shows a screenshot from when the site was functional in this way.

Concerns continue to persist about privacy and location-based information. Although the feature is not turned on by default, users can still share their FourSquare check-ins via Twitter. Facebook's default settings share all of users' updates (including check-ins) with everyone online.

We'll take a closer look at privacy in Chapter 16.

Conclusions

In this chapter, we looked at the relationship between location and social media. Users' location can be shared explicitly in profiles or through a mobile device's location service. It can also be estimated from a user's IP address. Users may choose to share their location to notify friends when they are out or to keep a log of their activities.

Locations can also be used by third parties to monitor larger phenomena such as outbreaks, crisis situations, or the spread of information across geographic areas.

Privacy is a major concern with location-based social media. The issues are similar to those relevant for most social media, but with the sensitive nature of location, they take on bigger importance.

[2] http://pleaserobme.com

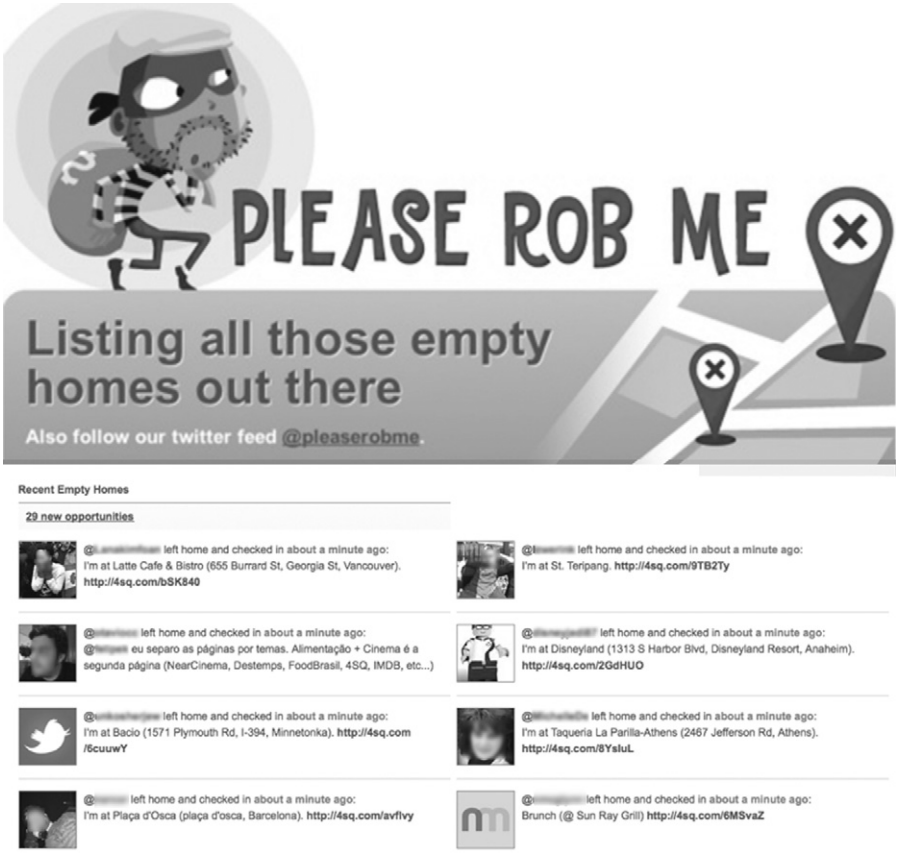


FIGURE 12.7

A screenshot of Please Rob Me, showing the Twitter identities of people who left home and checked in elsewhere. The site is no longer functional, but it illustrates what can be done with location information that is overshared.

Exercises

1. List three ways a social networking company, like Facebook, might make money using the location information provided by users.
2. List three offline social activities that could be enhanced by integrating people's location. Explain each example and describe the role of location data and how it helps.
3. List three online social activities that could be enhanced by integrating people's location. Explain each example and describe the role of location data and how it helps.

4. Imagine you had access to the location from where every Twitter message was posted. List three features you could add to Twitter that used this location data.
5. You are charged with designing an app for a mobile device, like a smart phone. Assume that you know the user's current location and the location of other people using the app. Design an app that has a social component and that leverages location information. Sketch the main screens of the app. Then describe the following features of your app:
   a. What is its purpose?
   b. What key features does it have?
   c. What is the social component?
   d. How is location used?
   e. How does location improve or relate to the social experience?
6. Imagine your favorite social media site released an update to its mobile application that uses your location. Consider location-based privacy issues with respect to the following uses:
   a. Imagine this is integrated into a new feature that allows you to pull up a screen that shows you any of your friends or followers who are physically nearby. It would similarly show those people that you are nearby. The list of friends or followers would link to each person's profile page.

      Would you use this app? Why or why not?
   b. Now imagine a different feature that allows you to pull up a screen that shows any users of the social media service which other users are nearby (i.e., it is not limited to your friends). This would allow you to know who is nearby, to see the profiles of any user in your vicinity, and to friend them more easily.

      Would you use this app? Why or why not?
   c. Compare the privacy concerns that arise between the apps in parts a and b. Which issues are the same? What issues are unique to each app?
7. This chapter presented examples of how location data and social media were used to track the flu, monitor wildfires, and respond to crisis events like earthquakes. Think about issues facing your local community and describe the way that location information and social media could be used together to help address those issues.
   a. Sketch out the main pages or screens of your application. This could be a website, a mobile application, or something else. Describe all of its main functionality through pictures and with text where necessary.
   b. Describe the role of social information.
   c. Describe the role of location information.
   d. Explain how the social and location information will be integrated. Will it come automatically from users' devices? Will they have to check in? Will volunteers do the integration manually? Describe any challenges that might arise when trying to integrate this information.

   e. What do you think would motivate people to use this system? Does it address a real problem they have, or is it creating a new opportunity for them to use technology for something they're currently addressing without it?
   f. What would keep people from using your system? Are there privacy concerns, technical challenges, technology access, or other issues that might drive people away from the application?
   g. If you were running a company, would you be able to make money off this application? If yes, how? If no, does that matter? Where would the resources come from to support the application then?

8. Download an application for your smart phone that uses location information. This could be a location-based game like FourSquare or Geocaching, a workout tracker like Run Keeper or MapMyRun, or some other application. Make sure that you can see the locations that you logged through either the application or its website.
   a. Use the application for one week.
   b. Make a report showing your location activities for the week. This may be a list of all the places you checked in, a collection of maps of your workouts, or a trace of your location activities. All the information should come from the app, and your report should summarize all the location data that the app collected.
   c. What is the social aspect of the app? How is your location data shared?

CHAPTER

Social Information Filtering 13

How does a user process all this information? Some of it will be more useful than the rest, and the sheer volume often means a filter would be helpful to sort through it all. This chapter introduces several methods used to sort, filter, and aggregate information from social media using social connections.

Social sharing and social filtering

One way to find useful information among all the links, news, videos, and photos posted each day is to rely on other people to find it for us. Social sharing and social filtering use the interests of others, especially friends on social networks, to highlight information that is more likely to be of interest.

Social-sharing websites, like Digg, Slashdot, and reddit, are designed for people to share interesting content. The community then votes items up or down, and the most interesting links are highlighted. The reliance on large numbers of people to help complete a task like this is a type of *crowdsourcing*. From the "crowd" of people online, each contributes a tiny amount of work by sharing or voting on content, and the aggregate results are a valuable contribution.

On social news sites like those mentioned above, users share information with the goal of reaching many other people, and often with the desire to have their shared content appear on the site's main page. When a user's interest is limited more to sharing with his or her friends, Facebook or Twitter may be the site where the link is shared. Some sites, including Facebook, are also tracking users' behavior to share links to content they are viewing, without the user taking any active action to share.

Frictionless sharing is a term coined by Facebook founder Mark Zuckerberg to describe this trend, which is present on Facebook as well as many news sites and web portals. This type of sharing is most often seen in applications called social readers. When a user activates a social reader, either through a social network or a news source, the application lists all the articles that the user clicks on. The user takes no action to share these clicks; instead, the application shares everything the user does. This is motivated by the idea that the things a person's friends view are likely to be of interest. This does raise privacy concerns for many users who do not want all of their activity shared, but the practice is particularly popular among content publishers, like news sites, who see it as a new way

to highlight interesting content and pull in readers. Aggregating behavior that may show the "top-viewed" articles without information about who viewed them is one way around the privacy issues.

Automated recommender systems

Recommender systems are major parts of e-commerce sites and social media sites. We introduce the major types here and discuss how they take advantage of social patterns and connections to suggest items that users might like.

Even if the term *recommender system* is not a familiar one, nearly all Internet users will be familiar with them. These are the features of websites that suggest items a user might like. Amazon.com uses one to suggest other items a customer might want to buy. Netflix uses it to suggest movies that a subscriber might want to check out. Pandora uses it to automatically generate Internet music channels that match a user's taste. All of these personalized suggestions based on a user's previous activity come from recommender systems. They rely on explicit data, such as user ratings, or implicitly captured data from users' behavior such as making purchases or viewing an item.

Because good recommendations keep customers interested in a website and increase the likelihood that they will buy something, they have become big business. As more and more user-generated content comes online, new algorithms for generating recommendations are created. These methods for generating suggestions are often quite complex, but the core idea behind many of them is to take data created by other people and personalize it for the individual user. It is an excellent example of aggregating information, especially ratings of items, in a social way.

Traditional recommender systems

Recommender systems basically work in one of two ways: suggesting items similar to the ones a person likes or suggesting items liked by people who are similar to the user. They might look at all the items that a user has rated and then look for items that are similar to the things the user likes. This is how Pandora, the online music streaming service, works. A user starts with a song or artist, and Pandora creates a musical profile of it. Then, Pandora selects songs that are similar in profile and plays those. If the user gives a new song a thumbs up, the profile of that song is combined with the existing profile to create a new set of attributes that the user likes. If the user gives a song a thumbs down, then the attributes of that song are deemphasized in the profile. This tactic of finding items similar to what the user is known to like is called item-based or model-based recommendation.

Table 13.1 Ratings by Three Users for Five Different Movies. Ratings are on a 1–5 Scale

	Star Wars	Jaws	Wizard of Oz	The Godfather	2001
Alice	5	4	3	3	1
Bob	3	5	2	5	1
Chuck	4	3	2	2	2

Item-based recommendation is not very social in that it does not rely on other people very much, but the second type of recommender systems relies entirely on other people's actions. These work by finding people who have similar tastes to the user and then recommending items that those people like. This is called *collaborative filtering*. At its core, collaborative filtering looks at each pair of users, finds the items that both people have rated, and computes a similarity score for the two people based on their ratings. That similarity measure is then used to give similar people more say in how much the user might like a new item.

Consider this simple example of collaborative filtering. A user, Alice, has rated a set of movies. Two other users, Bob and Chuck, have also rated those movies. These are shown in Table 13.1.

There are many ways to compute similarity with Alice. One option would be the average difference between ratings of these movies. In this example, Bob has an average difference of 1.2 with Alice, while Chuck has an average difference of 1.0. A more common measure of similarity is the correlation between the ratings. Chuck's ratings are 1 point lower than Alice's for every movie except *2001*, where it is 1 higher. His ratings track very closely with hers. Bob, on the other hand, does not seem to follow any pattern of being higher or lower with respect to Alice. This idea is captured by the Pearson Correlation Coefficient, a simple statistic that measures how well aligned two sets of values are. You can compute the Person correlation in most standard spreadsheet applications, including Microsoft Excel. It is always a number between -1 and 1, where a higher positive number indicates a high similarity and a negative number indicates preferences that vary in the opposite direction. In this example, the correlation between Alice and Bob is 0.26 and the correlation between Alice and Chuck is 0.83. Because correlation is commonly used in collaborative filtering, the rest of this example will use those values.

Now assume Alice wants to know how much she might like the movie *Vertigo*, which she has never seen. Both Bob and Chuck have seen it. Bob rated it a 3 and Chuck rated it a 5. What would be a good recommendation to Alice for how much she will like it? One option is to show the average rating for the movie, which is a 4 in this case. However, that does not take into account that Chuck is more similar to Alice than Bob is. A simple example of collaborative filtering will use the correlation numbers to compute a weighted average. Bob and Chuck's ratings will be multiplied by their correlation with Alice, and that total will be divided by the sum of the weights.

$$\frac{\underset{0.26*3}{\textit{Bob}} + \underset{0.83*5}{\textit{Chuck}}}{0.26 + 0.83} = 4.5$$

Notice that this weighted average comes out higher than the simple average. That is because Chuck gets more weight, and since he is more similar to Alice, his higher rating of the movie is given more consideration. Thus, the recommended rating of *Vertigo* for Alice is 4.5 stars.

The ratings produced by recommender systems can be used directly to indicate to a user how much they might like a particular item, or they can be used to sort items, showing those that seem most promising higher up in a list. They can also be used to filter out items a user is unlikely to like.

Social recommender systems

Collaborative filtering is an early example of how algorithms can leverage data from the crowd. Information from a lot of people online is collected and used to generate personalized suggestions for any user. These techniques were originally developed in the 1990s and early 2000s. Since the availability of this data has increased with the rise of social media, recommender systems have started to consider social connections in addition to similarity.

Simple examples of social recommendations can be found on many social networking websites. For example, on Twitter, when a user searches for a term, the search results can be shown in three ways: all tweets that match the search, "top" tweets, as determined by Twitter, or tweets only posted by people the user knows. This simple social filter excludes anything from unknown people, since it may be of less interest.

Friend recommenders are also common in social networking websites. Facebook prominently features a "People You May Know" section, which is essentially a recommendation of people to add as friends. This uses social network data to guess at what edges might be missing from a network. For example, if you are friends with 9 out of 10 densely connected people, it is likely that you are also friends with the 10th person.

Social relationships can also be used with collaborative filtering algorithms. The similarity measure that these algorithms traditionally use can be replaced with a variety of statistics taken from the network. Using trust or tie strength would give more weight to people who are close to the user and likely share similar opinions. Trust has been particularly well studied for making recommendations, and systems exist that leverage it for applications as diverse as recommending a movie to recommending mountain ski routes. An example of such a system will be described later in this chapter.

Case study: Reddit voting system

Reddit is a social news website. Users post links to interesting items, and other users can vote items up or down. These votes are used to rank each article, and

top articles appear on the site's front page. Users can comment on links and those comments can also be voted up or down.

These votes are extremely important for highlighting interesting and new content, but for this purpose, a simple vote count is usually flawed. It is biased toward older posts, since they have more time to gather votes, and they also stay highlighted longer, making them more likely to receive new votes. One possible solution would be to discount older votes, but that might unfairly discount good content that received a lot of up votes early.

Reddit has implemented an interesting system for counting votes, so older comments are treated fairly. Consider two posts. One was posted several days ago and has 45 up votes and 5 down votes. The second post was posted just a few hours ago, and has 5 up votes and 1 down vote. Counting votes clearly places the older article on top, since it has far more up votes than the newer article.

Another option is to consider the percentage of votes. The older article has a 90% up vote percentage (45/50), while the newer article has an 83% up vote rate (5/6). However, this might not be fair to the new article. Since it only has six total votes, the single down vote has a huge impact on the overall percentage score, while a single additional vote on the older article would have a small impact. Or consider an article with only one positive vote, which would have a 100% up vote rate, but clearly doesn't have enough data to go on. However, this is getting closer to the idea of treating articles equally regardless of age.

To eliminate the impact of time, imagine that both posts sat for months, accumulating votes, and *then* compare their percentage of up votes. That would be a fair comparison. Obviously, a system does not want to wait months before ranking items, but if it can estimate with high confidence the number of up and down votes a post will get, then that estimate could be used to compare articles regardless of age. This is the strategy that reddit uses forits "Best" sorting of comments.

Then look at the number of votes a post has received and use a statistical method to estimate the 95% confidence interval to forecast how many up votes a post will get. This is similar to how political polls work; they talk to a sample of people and find out how they would vote, and then they use their responses to extrapolate how everyone would vote.Posts with very few votes will not confidently have more up votes than an article with many votes, so they will stay at the bottom of the ranking. However, for an article with enough votes that a 95% confidence interval estimates it will receive more up votes than an older article, it will appear higher in the rankings. Although the 95% confidence does not always accurately predict the number or proportion of votes a post will receive, it adjusts after every vote so that something incorrectly placed high in the rankings can quickly be moved down.

This system is interesting both for how it works and for what it says about the importance of ranking in social-sharing systems. Reddit has several other ways of sorting comments ("top," "new," "old," and "controversial" are also options). The fact that there are so many sorting options indicates how important it is for users to sort through all the content. The effort the reddit community put into

developing these options, as well as the "Best" ranking described above, show the various ways they feel that social feedback—votes on user comments posted to user-shared links—can improve the way they present information.

Case study: Trust-based movie recommendations

As described above, traditional collaborative filtering recommender systems compute some measure of the similarity between two people. Those similarity measures are used to estimate how much a person will like an item by taking all the ratings for the item and giving more weight to ratings created by similar people. Social recommender systems may replace the similarity measure with social features, like trust or tie strength. This section introduces one example of a trust-based recommender system for movies.

FilmTrust is a research prototype of a movie recommender system that uses social networks and trust. The website allows users to rate and review movies, to view others ratings and reviews, and to make social connections with other users. These relationships do not have to be reciprocal; as on Twitter, a user can follow someone else if the user is interested in that person's movie reviews. When creating a social relationship on the site, users are required to rate how much they trust the other person's opinion about movies on a scale from 1 (very little trust) to 10 (very high trust).

With that data, the system knows how much a person trusts her friends. However, her friends likely haven't seen every movie she is interested in, and she will want to consider the opinions of other people. A typical collaborative filtering algorithm would consider the similarity between the user and each other person. Trust-based recommenders, like FilmTrust, use social network information instead. The system estimates how much the user might trust other people's opinions based on whom she trusts, whom her friends trust, and so on.

As a simple example, consider Figure 13.1. Alice wants to know how much to trust Dave, but she does not know him. There are many methods for estimating trust given a social network like this. For simplicity, we will present one example, which is the basic method used in the FilmTrust system.

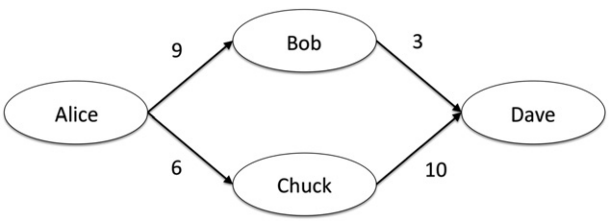


FIGURE 13.1

A sample social network with trust values between nodes.

Although Alice does not know Dave, she has two friends, Bob and Chuck, who know him. She can ask each of them how much she should trust Dave. Bob would say Dave is trustworthy at a level 3 out of 10, and Chuck gives Dave a 10 out of 10. Alice could average these values, and estimate Dave's trustworthiness at $(10 + 3)/2 = 6.5$. However, Alice trusts Bob's opinion far more than Chuck's. Thus, she might give more consideration to Bob's rating of Dave than to Chuck's rating. A simple way to do this is to do a weighted average. Alice has given Bob a higher trust rating, so his rating of Dave gets more weight. Let Trust(A,B) be the trust rating that Node A has for Node B. A weighted average would look like this:

$$\frac{Trust(Alice, Bob)^{*}Trust(Bob, Dave) + Trust(Alice, Chuck)^{*}Trust(Chuck, Dave)}{Trust(Alice, Bob) + Trust(Alice, Chuck)}$$

$$= \frac{9^{*}3 + 6^{*}10}{9 + 6} = 5.8$$

Notice that this weighted average of 5.8 is much lower than the simple average of 6.5. That is because Bob, who is trusted more, has a lower score for Dave. His low score gets more weight, so the whole inferred trust value from Alice to Dave goes down. Note that this trust value is personalized for each user. Chuck still trusts Dave at a level 10, and Bob still trusts Dave at a level 3. Alice's personalized trust estimate based on whom she trusts and whom they trust is a 5.8.

As mentioned above, there are many algorithms for inferring trust values in a social network like this, and this example is only one simple case. However, it illustrates how trust may be carried across paths in a network to come up with estimates. These algorithms can be used to compute an estimate of trust between any two nodes in the graph that are connected by a path.

The next step is to use the trust values, both direct ratings from the user and inferred values, to recommend movies. Each movie in the system will have a set of ratings, and the average rating is often shown in many systems. However, to do a recommendation, the system will compute a personalized estimate of how much it thinks a user will like the movie. The FilmTrust system does this in a simple way, replacing the similarity measure used in collaborative filtering with the personalized trust value for each user. The idea behind using trust is that the movie ratings of trusted people should be considered more strongly than ratings from less trusted people.

Continuing with the above example, say Bob, Chuck, and Dave all saw and rated the movie *Night of the Living Dead*. Say their ratings are as follows on a five-star scale:

- Bob: 5 stars
- Chuck: 2 stars
- Dave: 4 stars

Their average rating is 3.67 stars, but what is the personalized recommended rating for Alice? FilmTrust will do a weighted average, multiplying Alice's trust for each person with that person's rating of the movie, and dividing by the sum of the trust ratings. In this case the value would be as follows:

$$\frac{\begin{array}{ccc} \textit{Bob} & \textit{Chuck} & \textit{Dave} \\ 9*5 + & 6*2 + & 5.8*4 \end{array}}{9 + 6 + 5.8} = 3.86 \text{ stars}$$

In this case, the recommended rating is a bit higher than the average rating, but how well does this approach work overall? Researchers compared the recommended rating computed as described above with users' actual ratings of movies. The difference between the recommended rating and the actual rating was an estimate of error. Overall, they found that the trust-based recommendations performed as well as standard collaborative filtering algorithms. However, when the user's rating of a movie moved away from the average rating, indicating that the user's opinion was very unusual, the trust-based ratings greatly outperformed collaborative filtering algorithm. The conclusion is that trusted people are especially valuable in cases where a person cannot rely on the general opinion to reflect her own. For example, when someone generally dislikes movies in a popular genre (slapstick comedies), trust-based recommendations tend to accurately rate those movies lower than traditional recommenders. Similarly, for movies with a cult following, like *Napoleon Dynamite* or *Little Shop of Horrors*, a trust-based recommendation is more likely to give an accurate rating.

The FilmTrust system also uses trust values to sort reviews. When users write reviews for a movie, they all appear on the movie's page, as is common with review-based websites. However, as discussed earlier, a problem with those sites is finding the most relevant reviews among tens or possibly hundreds of items. FilmTrust computes the personalized trust value between the user and each review author, and shows the reviews from the most trusted people first. The motivation behind this choice is that trusted people will likely reflect the user's tastes best, and thus their reviews will be most relevant and should appear first.

Conclusions

By using social network analysis techniques to identify information about people, their role in the network, and their relationships can be useful to help sort, aggregate, and filter information in social media.

People can share content directly by sharing links on their social networking pages, posting them on social-sharing websites like reddit, or more passively sharing with social readers that show friends everything a person has looked at. These methods highlight information that a person's friends have found interesting, and that filter is often very useful for identifying good content. Further user

input, like votes up or down on content, can further help sort and filter information shared in this way.

Recommender systems move up a level, aggregating ratings or behavior and using that to personalize suggestions for items that a person might like. Collaborative filtering systems use similarity estimates to show items that people similar to the user like. Social recommender systems replace or enhance the similarity measures with social features, like trust relationships, to recommend items. This leverages social information in several ways, using ratings that users supply and their social connections to highlight interesting items.

Leveraging social information is already effective for finding information, and as the amount of information people encounter online grows, it will be important to develop new methods that incorporate social information and techniques for filtering, sorting, and aggregating content.

Exercises

1. Go to one of your social media pages—Facebook, Twitter, etc.
   - **a.** View all the posts from the last day, and rate them for importance on a 1−5 scale where 1 is unimportant and 5 is very important.
   - **b.** For each person who has posted something, list all the ratings you gave to their posts.
   - **c.** Do some people stand out as posting more important information than others?
   - **d.** If some people stand out as sharing more important information, are there any attributes (personal attributes or attributes of your relationship) of the people who appear more important that stand out? Are there attributes of people who post less important information? Is there a reason you follow people with less important content? Explain the pattern using social features.
   - **e.** If there is no pattern showing who shares the most interesting information, why do you think that is? Do all your social contacts have equivalent relationships? Have you already applied filters? Do they each contribute important things in different contexts? Using social features, explain why you think there is no pattern.
2. Look up four of your favorite songs from different time periods and artists on Pandora (item-based recommender system) and LastFM (collaborative filtering recommender system).
   - **a.** For each song, list five other recommended items.
   - **b.** How are the recommended songs similar or different from the two different types of recommender systems?
   - **c.** What accounts for the differences?

3. Look up your favorite movie on Amazon.com (find the DVD or Blu Ray version), Facebook, the Internet Movie Database, and Rotten Tomatoes.
   a. What information does each site provide in terms of reviews and ratings?
   b. What information does each site provide in terms of recommendations for other movies?
   c. Which sites, if any, use your social network to provide information? What information do they use?
   d. What information is most useful to you from everything available from all the sites? Why?
4. Imagine you have started a new movie website. You will have information about every film and you want to use social features to show readers the most relevant movie reviews.
   a. Describe how you will sort the reviews using social data?
   b. What social information will you need about the users, if any? How will you use it?
   c. What type of interaction will you require from users on the site, if any? For example, will they need to vote or can they simply log in and look at the site?
5. Find a social reader application for a major online information source (e.g., *The Washington Post*, Yahoo!, or The Onion).
   a. Sign up to use the application or connect it to your social media account.
   b. What articles does it recommend to you?
   c. How is it making those recommendations?
   d. What is it showing you about your friends' behavior, if anything?
   e. Read some articles through the social reader. What appears on your profile page within the application? What information, if any, appears on your social media profile page (e.g., your Facebook wall)?
   f. How useful are the recommended articles? Explain your answer.
6. Imagine you are an online information provider, like a newspaper or a web portal. Many of your users used their Facebook or Twitter accounts to log in on your website to keep track of their preferences. As a result, you have access to all of their social relationships and profile data. You can also post to their profile page (e.g., Facebook wall or Twitter timeline).
   a. Create a part of your website that uses this social information. Explain what you will do, what social information you will use, and sketch out the major screens that would appear on your site.
   b. How did you use social information to improve the way the user accesses information on your website?
   c. Add a feature that will help draw more people to your website. Explain that feature, how it is using social information, and why you think it will work.
   d. What are the privacy concerns, if any, that arise from your technique? How would you address those concerns?

7. Imagine you want to use social information to filter your email.
   a. Which messages would you want to appear first?
   b. Using Gephi or NodeXL, create your email network.
   c. Which people appear to be most important in that network? Which measures did you use to determine that?
   d. What social network characteristics would define the people whose messages you think are most important? These could be structural network statistics or relationship-based information (like trust or tie strength).
   e. How would your email program compute the characteristics described in (d)? What factors should it consider?
   f. Look at the last 50 email messages you received or that remain in your inbox. Rank them in terms of importance. You could use a scale (e.g., 1−10) or a simpler ranking (e.g., high, medium, low).
   g. Compute as many of the characteristics described in (e) as you can. Use those to rank your email messages in terms of importance according to the statistics.
   h. How well does your social network ranking match up with the ranking you did in (f)? Can you think of ways to improve the ranking based on these results?

This page intentionally left blank

CHAPTER

Social Media in the Public Sector

14

Traditional media, like newspapers, radio, and television, are effective ways to reach large audiences. However, traditional media is broadcast—messages go out, but there is little interaction with the audience. Information cannot easily be personalized, and quick, real-time responsiveness to audiences is difficult or impossible in most traditional media. Social media, on the other hand, is based around interaction. It allows personal communication, personalization of broadcast messages, and it can be updated quickly to keep people apprised of a rapidly changing situation. One could say that traditional media is one-to-many communication, while social media is many-to-many. These benefits are making social media increasingly popular as a communication mechanism in the public sector—among government, politicians, emergency-response officials, and others. This chapter will introduce analysis questions important to the use of social media in the public sector and present several related case studies.

Analyzing public-sector social media

When analyzing pubic sector use of social media, it can be approached as an analysis of (1) how public-sector users are taking advantage of the technology or (2) how people are talking about public-sector-related topics. Based on the approach, the major types of questions vary. This section will present guidelines on how to conduct both types of analysis.

Analyzing individual users

Public-sector users may be individuals, such as elected officials; local organizations, such asschools or libraries; or government agencies.

A social media user, whether an individual or organization, chooses to use social media for many reasons, but there are three major types of use:

- Broadcast/Sending information—A user may want to share information with followers or friends. These may be updates, requests, or other information. This use leverages social media because the audience is presumably interested in what the user has to say. These broadcast messages may be used to simply send information to followers or to engage in one of the following interaction types.

- Request Feedback/Input—Social media audiences may be willing to share information, whether it is an opinion about an issue, information they have about a crime, or posting the location of an event. A public-sector users may want to gather this information from its audience and thus may request the feedback or input in their posts. They may also utilize social media as an open channel for people to send them comments.
- Conversation Interaction—Unlike a request for feedback, this type of use encourages conversation or interaction directly between the user and individual members of the social media audience. This may be an elected official having a conversation with constituents or a librarian talking to a patron through social media to help recommend books. Unlike requests for feedback or input, where an audience member's input is simply accepted and processed, this type of interaction supports back-and-forth communication between the user and audience member.

There are many ways to analyze how an individual organization is using social media. Based on your particular interest, you may create your own hypotheses and questions, but the following are general guiding questions.

- Who is doing the posting?
- Who are the target audience members?
- Why is the audience engaged in social media with the organization? What type of content or interaction is the audience interested in?
- What are the goals of the user? Which of the three interaction methods above are they using?
- How is the user using social media?
- Do the user's actions support the goals?

These questions and guidelines should serve as a starting place for analyzing social media usage. Depending on the user and analysis, more specific questions and hypotheses are likely to arise and will lead to deeper analysis. To apply these questions, we will now look at several case studies of social media usage in the public sector.

Case study: Social media to solve an attempted child abduction

Just before 4 P.M. on July 17 2012, a 10-year-old girl and her 2-year-old brother were walking home from buying flavored ice in their South Philadelphia neighborhood. A man in a white car followed them for several blocks before parking and approaching the pair from behind. He grabbed the girl, covering her mouth and carrying away from her brother. The girl fought back, kicking and biting her attacker, and broke free of the man's grip. He dropped her and then quickly fled the scene.

FIGURE 14.1

A scene from the surveillance video released by the Philadelphia Police Department on YouTube and through other social media to help capture the man who attempted to abduct a 10-year-old girl.

The children did not know the abductor, but the incident was captured on several surveillance videos. The Philadelphia Police Department's Special Victims Unit released the videos to the public less than a day after the event occurred. Figure 14.1 shows a still from one of those videos. The police immediately started receiving tips via social media channels. Within hours of posting the video, the man turned himself in, claiming that he "felt that he could not walk, talk or breathe out there," according to Philadelphia Police.

The Philadelphia Police Department has been a pioneering user of social media, actively using YouTube, Facebook, and Twitter to gather information about crimes. They also have smart phone apps that let people report incidents and find local police stations. At the time of this abduction, the department reported catching 87 suspects through social media usage. In February 2012, they caught the abductor and rapist of a 6-year-old girl within 16 minutes of posting the suspect's photo. In another case, a suspected murderer was turned in by his mother after she saw his photo.

To analyze the Philadelphia Police's use of social media, the guiding analysis questions can provide help. In terms of use, they are taking advantage of social media both to broadcast to a large audience and to receive input from their audience. They are generally not having conversations in the social media, since the nature of their work means it is often better for a police officer to contact a person directly and privately to have an extended interaction.

Looking at the YouTube, Twitter, and Facebook accounts of the Philadelphia Police, we see that they are generally posting crime alerts and videos, to redirect other social media users to 911 dispatches if they are having an emergency, and to occasionally share departmental news. To be effective, they must not share too much information such that audience members would be overwhelmed, or information that is not relevant to them. The Philadelphia Police post to twitter usually

less than 10 times a day, ensuring that they do not overwhelm users with too much content. But is social media also an effective way to receive information from people?

Another case from Philadelphia holds some clues. A man was attacked on a city bus, but no one on the bus was willing to assist him during the attack or call 911. However, after police posted the video of the attack online, several witnesses identified the suspect. They were more willing to report the event electronically than they were to get involved at the time.

A full analysis would look at a more fine-grained breakdown of the types of content being posted and the frequency. It would also look at the people who follow the Philadelphia Police. Demographic information, like location, may suggest some reasons they are interested in the content. Surveying people about why they follow the department may also provide more insight into how the Philadelphia Police are effective, what they might do differently or better, and how other departments might interact online to be similarly effective.

By understanding the attributes of the social media users and the interactions between police and users, the police can optimize their social media strategy. Knowing the kind of people who are involved in an organization's social media and knowing their interests allows an organization to post content that keeps the audience engaged, that takes advantage of their knowledge, and that thus allows the organization to more effectively get its message out.

The Philadelphia Police is only one example of how police and emergency response officials may use social media. Social media also allows people to report information about emergencies on location. As discussed in Chapter 12, Location-Based Social Interaction, mobile social applications are allowing people to report locations of wild fires. Systems have been proposed that leverage social media in communities to share information about ongoing incidents (Wu et al., 2008).

Case study: Congressional use of twitter

In mid-2012, nearly 400 members of the U.S. Congress had Twitter accounts. Although all these congresspeople represent constituents and have similar duties and goals in their positions, they use Twitter in very different ways. Some have staff members post to their accounts, generally sharing links to press releases and other official information. Other congresspeople tend to their accounts themselves, carrying on lively conversation with their constituents about issues, bills under consideration, and current events. While analysis of each account will reveal these differences in usage, analyzing all members of Congress together paints an interesting picture of the type of content being shared by these similar users. This case study is a summary of work first presented in (Golbeck, Grimes, and Rogers, 2010).

Many insights are available by analyzing the way a specific group uses social media. The first step is to understand the needs of the group. What type of information do they want to convey? What type of interaction do they want to have with users? What type of activities do they undertake offline that might be well supported by social media? Then, analyzing how they are actually using social media may reveal a number of insights. Their existing utilization can be compared to their needs. Are they using it in ways that meet those needs? Are there needs that are not being met? Is social media being used to create new types of interaction? Could social media be leveraged to accomplish tasks for which it is not being leveraged?

Researchers collected over 6,000 tweets posted by members of the U.S. Congress over a six-month period. They read each tweet and categorized it into one or more of the following categories:

- Direct Communication—a message directed at a specific person either with the @id convention or in the text of the message. Direct Communication was divided into two mutually exclusive subclasses.
  - Internal Communication—This included messages from one congressperson to another or from a ongressperson to a staff member.
  - External Communication—All other messages, such as those to constituents, were marked as external communication.
- Personal Message—These are nonbusiness-oriented messages or notes, such as holiday greetings or other personal sentiments.
- Activities—A message reporting on the congressperson's activities was divided into two mutually exclusive subclasses.
  - Official Business: This included any official business in Congress, including voting, committee meetings, or making speeches on the house floor.

    Example Tweet: keithellison: Marking up the Credit Cardholder's Bill of Rights in the Financial Services Committee
  - Location or Activity: This code was used when a Congressperson was describing non-official activities including trips, meetings with constituents, lobbyists, or non-Congressional organizations, or activities in the home district.

    Example Tweet: neilabercrombie: @neilabercrombie just completed weightlifting workout at the Nuuanu Y. Advertiser featuring him on July 10; it's part of a regular feature.
- Information—This code describes a message that provides a fact, opinion, link to an article, position on an issue, or resource.

  Example Tweet: greshambarrett: Barrett announces 10 campaign events: Congressman Barrett will campaign in his district this week, http://tinyurl.com/6puxze.
- Requesting Action—When a congressperson requests constituents to take some action like signing a petition or voting, the message is coded this way.

- Fundraising—Messages occasionally ask for donations and contributions, and we code those as fundraising.
- Unknown—Some messages cannot be classified, as when they are only URLs with no text, test messages, or other mistakes such as a single character.

This is an example of how a finer-grained content analysis works. Researchers develop these types of categories by going over their data several times (in this case, reading the tweets) and coming up with a coherent and comprehensive set of categories to describe the content. Then, the data is reviewed again, placing each item in the appropriate category. This process is called *coding*. Ideally, two or more independent researchers code the data. Then, their agreement is calculated (called *inter-rater reliability*). A high inter-rater reliability indicates an accurate coding of the data.

After each tweet was categorized, patterns emerged about how these congressional users were taking advantage of Twitter. The vast majority of posts were Information (sharing links, opinions, or facts) or Locations and Activities (posts about unofficial activities the congressperson was doing).Perhaps surprisingly, the members of Congress did almost no political fundraising over Twitter, despite the fact that the collected tweets covered an election cycle. Figure 14.2 shows the breakdown of tweet types that the researchers found.

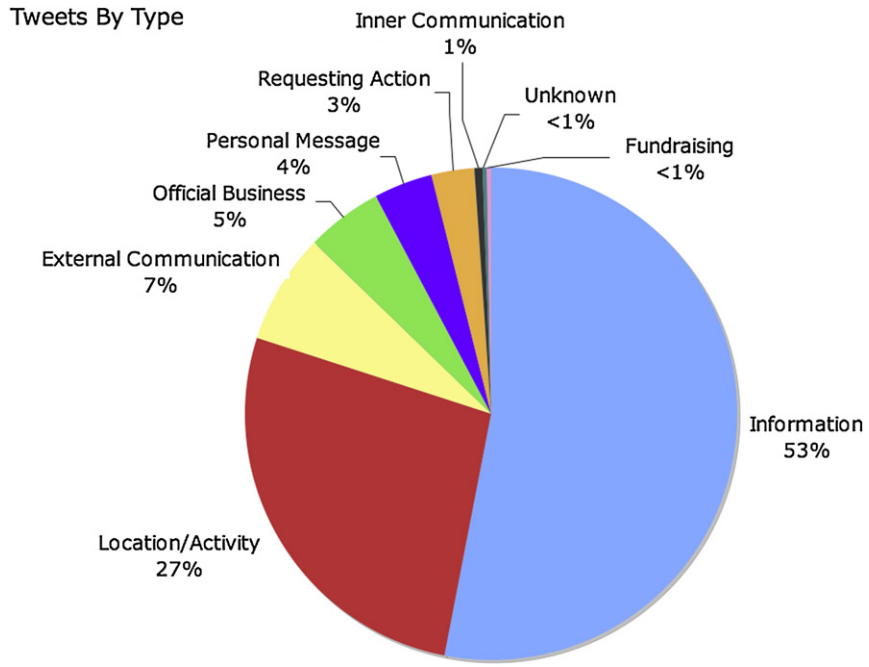


FIGURE 14.2

Types of tweets posted by members of the U.S. Congress as found in Golbeck, Grimes, and Rogers (2010).

Of course, as social media becomes more popular and widely understood, and as politicians and campaigns become more savvy with the technology, the patterns of use are likely to change. This case study serves as an example of how analysis of a group of users can be performed; it illustrates the type of activities common to the group, even when there are differences between individuals.

Case study: Predicting elections and astroturfing

The examples above have analyzed individual accounts or groups of individual accounts. This case study flips the perspective, analyzing behavior from a large and diverse set of social media users to understand public opinions about public issues.

While the study of how elected members of the U.S. Congress used Twitter showed little effort to raise campaign funds, social media has become increasingly important for political campaigns. The 2008 election of Barak Obama is largely cited as the first time a campaign truly embraced and used social media to reach voters. The power of social media to allow personal access to interested voters is powerful for campaigns, and the trends of how users discuss a campaign can provide valuable insight into what the public is thinking about an issue or an election. As a result, both individuals and the media have started using trends on social media as a way of understanding public opinion.

This has led some people to think that elections can be predicted based on how often a candidate is mentioned in social media and how positively or negatively that candidate is discussed. On the surface, this seems to be a valid approach. If social media is full of positive comments about Candidate X and there are fewer good posts about Candidate Y, then it would appear that Candidate X has more support and thus is more likely to win an election. There was also some anecdotal support for this technique, including the 2009 German elections (Tumasjan et al., 2010) and in the 2010 U.S. congressional elections (Livne et al., 2010).

However, further studies showed that using social media had only a slightly-better-than-chance success rate at predicting elections (Metaxas, Mustafaraj, and Gayo-Avello, 2011). Furthermore, the volume of social media posts about a candidate were not necessarily representative of the public's opinion or conversation overall. A vocal minority could often overwhelm a silent majority, as was observed on Twitter in a 2010 special election for a Massachusetts Senate seat (Mustafaraj et al., 2011).

Although the trends and popular topics on social media may not reflect the public's opinion in general, it is still powerful to see a lot of discussion about an issue, particularly if it is favoring one side. Grassroots efforts often utilize social media to build interest in their causes and rally support. Since social media is inexpensive to use and can reach a large audience, it can be a very effective tool

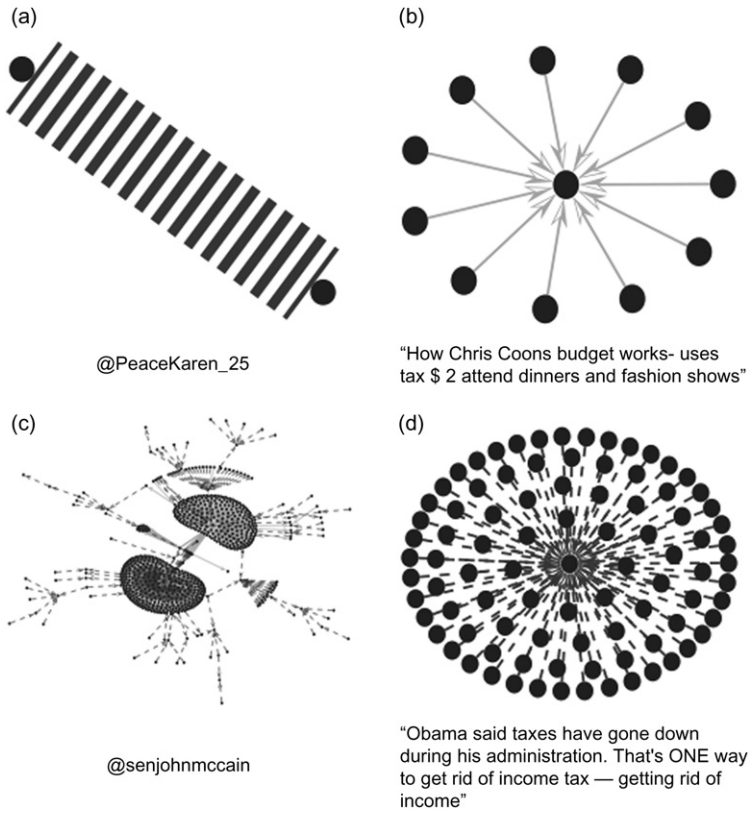


FIGURE 14.3

Examples of Twitter behavior from Ratkiewicz et al. (2011). Dark edges indicate re-tweets and light edges indicate mentions. Graphs (a) and (b) are astroturfing accounts, while graphs (c) and (d) are real accounts.

for gathering support and drawing attention to an issue. The success of grassroots movements online and the attention people are willing to pay to these efforts have also caught the attention of larger organizations. Since messages coming from large companies or political organizations may not garner the trust that a true grass-roots effort might receive, the large organizations have sometimes resorted to creating fake grassroots campaigns. This strategy is often called *astroturfing* (for its fake grassroots).

The technique is not unique to social media. Politicians and public officials have a long history of sending "letters to the editor," written under false names, in which they attacked their opponents or advocated for their own policies. Within social media, astroturfing is relatively common. Researchers have developed a tool called Truthy to detect astroturfing content on Twitter (Ratkiewicz

et al., 2011). They present several examples of fake accounts set up on behalf of politicians detected by their system. The analysis includes looking at the social network connections between accounts. For example, Figure 14.3 is taken from their work. It shows two examples of fake accounts and the topics they discussed, together with two examples of real accounts. The difference is in the structure of the social network.

Media investigations have also revealed a deep system of astroturfing. A recent report[1] shows that PR firms have "persona management" software that they can use to create an army of fake accounts, prevent them from contaminating one another, and creating suspicious behavior patterns (like those shown in Figure 14.3). The software can automatically create posts and manage accounts, so that a few people can generate many posts from many accounts. This makes it look like there is a large grassroots movement for a position that is actually advocated for by a large organization.

Analyzing networks to detect legitimate trends versus organized, false accounts is complex. It requires an analysis of social network connections, content analysis, and some detection of the sources behind each account. This type of analysis is not simple, but it provides valuable insights into public opinion and efforts to shape it.

Exercises

1. Find an elected official who represents you and is using some form of social media.
   a. What types of interaction is he or she using (broadcast, input, communication)?
   b. What appear to be his or her goals with social media?
   c. Analyzing the account, does your representative appear to be effective with social media? Why or why not?
2. Choose a popular current issue of public debate (a bill under consideration, an election, or a political issue). Search Twitter for posts about that issue.
   a. What opinions are you able to find? Summarize them.
   b. Is one opinion dominating the others?
   c. Do you find a lot of content repeated? Perhaps one or two tweets that are repeated by many accounts? Does this appear suspicious, or is there a reason for it?
3. Look in the online galleries of NodeXL or ManyEyes. These websites include many networks of political or government agencies. Select two networks that represent graphs of similar entities. For example, you might choose two graphs of networks of politician's Facebook networks, two graphs of

[1] http://www.dailykos.com/story/2011/02/16/945768/-UPDATED-The-HB-Gary-Email-That-Should-Concern-Us-All accessed July 2012.

government agencies' Twitter networks, etc. For the two networks you have chosen, compare the graphs. Look at network statistics like size, density, clusters, communities, etc. Choose at least 3 points of comparison and explain what they mean for each of your chosen networks.

4. Choose a federal, state, and local public organization. This could be a library, elected representative, agency, or the like.
   a. What social media presence does each organization have?
   b. For each social media account for each organization, list how many people follow it.
   c. For each Twitter, Flickr, or YouTube account that each organization has, use NodeXL to create a 1.5 egocentric network. Compare the networks across organizations and account type. Using features such as density, clustering, and size, compare and contrast the networks.
   d. Create one recommendation for each of the agencies you chose about how they could improve the way they are currently using social media. Explain your recommendations.
   e. Create one recommendation for each of the agencies you chose that suggests new social media they could use to support their mission. Explain your recommendations.
5. We have seen several examples of how social media can be used in crises. Create a recommendation for a public agency (at any level—federal, state, or local) about how they could leverage social media in crisis situations. Describe your recommendation, why you think it would work, and how the agency should go about implementing it.

CHAPTER

Business Use of Social Media

15

The previous chapter described many ways the public sector could use social media. Businesses and nonprofit organizations can use social media in similar ways. They can use it to broadcast messages to customers, to solicit feedback or input, or to interact with customers. For businesses, there are often two major drivers of social media use—marketing and customer service—and these will often dictate how best to use social media.

Businesses can communicate with their existing customers, potential customers, or interested parties. For ease of discussion in this chapter, we will refer to all of these audience members as customers.

Social media can be used to reach many customers, and because they are an interested audience, it is an ideal way to share specials or updates. It is also a medium that encourages interaction, and can help reach customers with problems and to form more personal relationships. Monitoring what people are saying about a company is also important, and it can be integrated with the strategies mentioned above. Indeed, some have called social media the "world's largest focus group." Many tools are available that will monitor social media for any mentions of a company. They will include names of products, variations in spelling on the company's name (and nicknames), and even the sentiment (positive or negative) of the things people are saying. Additionally, comparative analysis can be performed to compare how people are discussing a company's product and competitor products.

In this chapter, we will look at ways of measuring how a business is using social media and how effective they are. We will also discuss how social network analysis techniques covered in this book can be used to gauge a business's success. Then, we will look at several examples of how businesses are using social media successfully to broadcast, monitor, h and interact with users, and what happens when those efforts go awry.

Measuring success

A common measure of effectiveness for business is return on investment, or ROI. This number is simply computed as

$$(\text{Income} - \text{Cost})/\text{Cost}$$

Essentially, it measures the fraction or percentage of income earned beyond what was spent. A campaign that costs $1,000 and that earns $1,500 for a company would show a 50% ROI (the company earned back 50% more than it spent).

While this is measurable in some cases, it is not always an easy statistic to compute with social media. Advertising campaigns that combine traditional media with social media may lead to increased sales and show a lot of engagement online, but it is difficult to measure how many of the sales come directly from social media activities. Similarly, if a company begins addressing customer service issues online, that may reduce the number of issues coming in through more traditional channels, like phone calls to a customer service line. However, it is not clear how this impacts income or expenses since people are required to offer service over both channels.

There are other ways to measure the success a company is having in social media. Social media success does *not* always mean business success, but the following measures can indicate a social media campaign's success.

1. Counts—Counting activity is usually quite easy in social media; often the numbers are displayed publicly by the social media site. This may include number of fans, followers, or friends to see the number of people engaged. It may also be number of views on a shared video, number of "Likes" on a post (or similar indications of favorable opinions), or similar counts of people viewing and appreciating content.
2. Social Sharing—The counts mentioned in #1 are counting the number of people or their personal actions related to a business's social media site. Sharing is even more important. This could be measured as the number of times an item that the business has posted is shared, the number of times it is mentioned or retweeted on Twitter, or similar counts of sharing behavior.
3. Engagement Rate—Counts of people and shares can both be useful, but if a business with one thousand fans gets the same number of shares or likes as a business with one million fans, it indicates that the smaller business is being more successful. Thus, computing the number of engagement activities (likes, shares, etc.) divided by the number of friends, followers, or fans will show how engaged the social media audience is with a business.
4. Interaction—For businesses interested in engaging with customers in social media, measuring interactions can be helpful. Counts of the number of customers with whom the business has engaged, the number of conversations, how long each conversation lasts, and how well the interactions are resolved will all indicate how well the business is doing.
5. Referral Rates—Often, businesses will use social media to drive people to their websites that are not part of the social media site. Counting click-throughs, which can be easily measured in server logs or with website analytic software (e.g., Google Analytics), can indicate how much traffic a social media site is driving to the business.

6. Importance and Influence of Users—As has been discussed in many places in this text, users vary in their influence and importance in social networks. For businesses, all the measures above treat users identically. In fact, having content shared by more influential users has a much greater impact. Thus, measuring the influence of users can be important. This can be done by computing centrality, if possible, or looking at simpler metrics like number of friends.

Influence is an ideal way to apply many of the social network analysis techniques covered in this book. As an example of these techniques, consider the Twitter network of @frontpageva, a restaurant in Arlington, Virginia. This is a small business with roughly 1,400 followers. Although their account is small compared to those of major corporations, @frontpageva actively uses Twitter to share specials and events and to interact with followers. Social network analysis allows us to see which of their users are most influential and what their reach is.

Beyond the number of followers, we can look at the engagement with @frontpageva on Twitter. Over the course of a summer month when hockey is not in season, they average between 350 and 400 mentions. Dividing that by the number of followers gives an engagement rate of around 30%, which is quite high. They also have roughly the same number of outgoing messages to other people online, indicating that they engage online with others as much as people send messages to them.

Reach is also important, especially for a small business. Figure 15.1 shows the 1.5 egocentric network of @frontpageva on Twitter. Each node is a follower of their account, and size and color indicate the number of followers of each person. Larger, lighter nodes have more followers.

Two clusters are apparent in this visualization. Diving deeper to inspect the nodes in each cluster reveals that the top group contains many Twitter accounts that share information about Washington, D.C. and Arlington, Virginia events, parties, and locations. The lower cluster is made up largely of Washington Capitals fans and players because the restaurant is across the street from the Capitals' practice facility.

The large white node in the lower cluster is John Carlson, a defenseman for the Capitals. Because he is a professional hockey player, he has tens of thousands of followers. In the upper cluster, there are a number of large accounts—though none as large as John Carlson's node. These represent popular venues and events in the Arlington area.

The @frontpageva account recognizes the importance of these followers. They regularly tweet to John Carlson, and those messages are visible to most of the Capitals fans in the lower cluster since they follow both accounts. John Carlson will also tweet to Front Page, and this online relationship provides valuable exposure and endorsements for the restaurant to their hockey-fan customers.

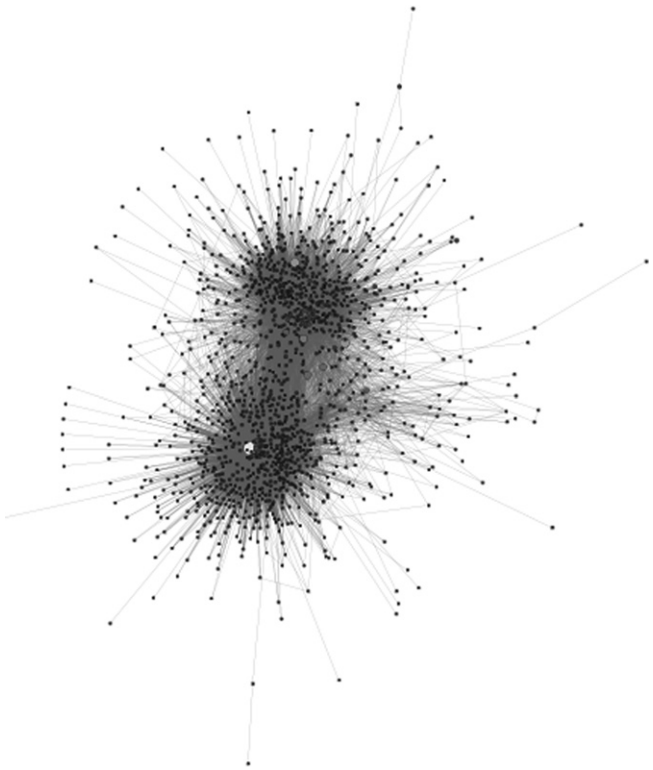


FIGURE 15.1

The 1.5 egocentric network of the @frontpageva account. Larger, lighter nodes have more followers.

This example illustrates how many of these metrics can be applied to analyze a business using social media. The next sections will describe examples of excellent (or poor) use of social media by businesses for specific purposes.

Broadcast example: Will it Blend? Marketing campaign

Blendtec is a manufacturer of high-end blenders. This may not seem like a product that lends itself to highly successful social media campaigns, but the company has one of the most measurably successful efforts of any business.

The company regularly releases humorous videos on YouTube in a series they call "Will It Blend?"[1] In the series, the company's founder demonstrates the

[1]Videos are viewable at http://www.youtube.com/Blendtec.

FIGURE 15.2

An example of a "Will It Blend?" YouTube video, showing the blender being used on an iPhone.

blender blending unusual items. These have included butane lighters, a skeleton, Justin Bieber memorabilia, and various electronics. Figure 15.2 shows an example.

The videos are not advertised in traditional ways and spread only by viral sharing, yet they have become extremely popular on YouTube. Their top-viewed videos have well over 10 million views each, and their collection of videos all together have 200 million views. The company's YouTube channel has over 400,000 subscribers.

As mentioned above, counting views or subscribers alone does not necessarily indicate that a social media campaign will help a business. In this case, Blendtec reports a 700% increase in retail sales since launching its YouTube effort.[2]

The videos increased sales by increasing recognition of the brand name and demonstrating the quality of the product. But why do people watch the videos? They are not high-quality productions, they do not feature famous people, and they are not advertised in traditional media.

Blendtec uses several strategies to draw attention. First, the videos show blenders blending things that pique the audience's curiosity; marbles, garden rakes, or guns are not things most people would blend at home. Second, Blendtec tags its videos well and blends items—like iPhones—that will be interesting to fans of the product being blended. This increases the chance for more views and shares from people who are interested in the items being blended.

[2]www.socialens.com/wp-content/uploads/2009/04/20090127_case_blendtec11.pdf, Accessed July 2012.

In general, Blendtec is successful because people share their videos, and many of them spread virally to millions of people. They get that attention by producing high-quality content that people naturally want to share, and it has had very good results for their business.

Interaction and monitoring example: Zappos customer service

Zappos is an online retailer that sells shoes, clothes, and accessories. It was actively engaged in social media, and in 2009 was named as having the best use of social media by Abrams Research, a company that focuses on social media strategy. Zappos has also been frequently cited for its excellent customer service, and the retailer tries to integrate this reputation into the social media.

Zappos interacts with customers on both Facebook and Twitter. On Twitter, Zappos maintains an account, @zappos_service, that answers customer questions and concerns. Representatives for Zappos who monitor the Twitter feed introduce themselves as they change shifts every few hours. A study by STELLAService, a company that studies online customer service, rated Zappos their best performer.[3] Over a 45-day period, Zappos was one of only two companies to respond to every service request within 24 hours, and the average response time was under an hour.

Zappos has hundreds of people in its customer service department, but only around 20 handle Twitter requests, and those are in short shifts every day. Usually there are one or two Zappos employees on Twitter at a time, so a very small fraction of the customer service team is needed to manage these requests.

An additional impact of offering customer service online is that other social media users "overhear" these interactions. They can see the back-and-forth conversations, and that helps them get an impression of the company's service.

Zappos serves as an example of how businesses can take advantage of social media to interact with customers. The Zappos service account not only responds to requests sent directly by customers, but it also monitors any posts about Zappos and sends messages to customers with concerns, even if those customers would not have contacted customer support to help with their problem. As a result, upset customers can be reached and helped, even if they would never have asked for help.

[3] http://happycustomer.stellaservice.com/2012/05/30/most-customer-service-tweets-go-unanswered-within-24-hours/ Accessed August 2012.

Social media failure example: Celeb boutique and the NRA

Measuring success is important, but social media can also have significant impacts on a business's reputation—both positive and negative. Social media gaffes are not rare, but businesses want to be extremely careful about them. Particularly because of the interactive foundation of social media, mistakes can be widely shared and backlash can come quickly.

This was evident on July 20, 2012. On that date at a midnight movie showing, an armed gunman opened fire in a crowded theater in Aurora, Colorado, killing 12 people and wounding dozens more. It was one of the worst mass shootings in history and became the top news item all night and into the following morning.

At 9 A.M., the National Rifle Association(NRA) posted this tweet from one of its most popular accounts:

> @NRA_Rifleman: "Good morning, shooters. Weekend plans? Happy Friday!"

Later that afternoon, an online clothing retailer, Celeb Boutique, posted the following tweet:

> @CelebBoutique: "#Aurora is trending clearly about our Kim K inspired #Aurora dress ;)"

Twitter users were outraged at both posts, inundating the accounts with negative comments. The NRA deleted their tweet three hours later, and eventually apologized and explained that it was posted by someone who had not yet read the morning news. Although this explanation was widely accepted, since the shooting had taken place in the middle of the night on the East Coast, the NRA took another step to mend the situation. Three hours after deleting the post, it deleted the @NRA_Rifleman account entirely, losing its 16,000 followers in the action.

Celeb Boutique also removed its tweet after an hour and then issued an explanation that it had not been aware of the shooting when it posted the tweet. This explanation was met with much more skepticism than the NRA's explanation. In addition to the tweet coming much later after the shooting, identifying the reason behind trending topics involves clicking only once on the topic to see the tweets about it. Furthermore, the company is generally a very savvy social media user, interacting personally and well with many people who mention them on Twitter. The "wink" emoticon at the end also led many to believe that they were trying to make an edgy joke. Unlike the NRA, after its initial apology, Celeb Boutique went on to resume its regular tweeting behavior.

What lessons can be drawn from these mistakes? The NRA example basically serves to emphasize that companies should be careful about what they post on social media because inattention to detail can produce a very costly backlash.

The Celeb Boutique case is more complex. The company was taking advantage of a "trending" topic; Twitter identifies the 10 most common words, phrases, or hash tags and marks them trending. Then, anyone can click on those trending

terms to see all the tweets about it. Celeb Boutique—and many other people and organizations—try to include trending content in their tweets to appear when people look at the posts about a popular topic. In this case, "Aurora" was trending because it was the name of the town where the shooting occurred, and Celeb Boutique took advantage of that to market a dress. While connecting a product to popular terms or ideas is often effective in generating traffic—as is the case with blending an iPhone in the "Will It Blend?" example—connecting a product to a negative idea on social media can generate very negative feelings about a brand.

Conclusions

Social media can be a powerful tool for businesses, but because it reaches so many people, companies must be careful about their posts since mistakes propagate quickly. Social media can be used to broadcast out to users, as in the case of the "Will It Blend?" campaign, to communicate with users, as with Zappos's customer service, or a blend of techniques to receive input and feedback. There are many metrics for measuring success, from simple counts of followers and engagements to measuring ROI. Many social network analysis techniques covered in this text can be used to identify people of influence within a network, and to guide strategy for reaching out to certain users.

Exercises

1. Come up with five companies or brands you interact with regularly. For example, the companies could be a beverage bottler, restaurant, clothing brand, or technology company. For each of the five, find all the social media accounts you can. These will usually include a Facebook page, often a Twitter or YouTube account, and they may be present in many other types of social media.
   a. List each company and their social media accounts.
   b. Find as many counts for each social media account as described in the section on measuring success.
   c. How often does the company interact on their social network site? Is it many times a day, a few times a week, or never?
   d. What kind of interaction is the company doing? Broadcast, request for input, direct interaction, or a combination? Provide an example of each.
   e. Assess the company's social media strategy. What are they doing well and why? What could they do better, why would that be better, and how should they do it?
2. Find a major company offering customer service on Twitter. Search Twitter to find the 10 most recent customer service interactions they have had.

   a. Was the customer service inquiry resolved?
   b. How many messages, on average, were sent between the company and the person sending the request? Did the company ever direct the conversation to direct messages?
   c. Was the customer service request handled by one single person all the way through, was it handled by several people, or is it impossible to tell?
   d. On average, how long did it take for the person's request to receive its initial reply from the company?
   e. Overall, would you say the company successfully handled the requests? Why or why not?
3. Find a company that has undertaken a viral marketing campaign over social media.
   a. What is the essence of the campaign?
   b. What metrics can you use to measure it (number of views, fans, likes, etc.)?
   c. Is the campaign ongoing, or did it run for a fixed amount of time?
   d. Are any statistics available to indicate the success of the campaign? If so, what are they?
4. Find two competing companies that both maintain Twitter accounts. Use NodeXL to create 1.5 egocentric network graphs of their networks. Compare the networks, considering features such as density, clustering, and size. List at least three major features and explain the similarities and differences between the networks.

This page intentionally left blank

CHAPTER

Privacy 16

One of the major challenges to using social media is privacy. Having access to so much personal data is a powerful thing. But although many users would not want their personal information used in this way, they are either unaware of the implications of sharing, or they do not have the skills to protect themselves. This chapter introduces some of the major privacy issues online and discusses ways to protect social information.

There are two major areas to consider in relation to privacy in social media:

how information is shared with other *social media users*, and
how social media websites and services distribute users' information to other *parties*

The main purpose of social media is to share information with other users, but people often want to control who sees what they post. Some information (like reviews, ratings, and comments) is not sensitive or especially personal in nature, and is thus often unrestricted and accessible to anyone. Other information (like photos, personal messages to friends, and contact information) reveals more about people, their relationships, and potentially intimate details. Users often want to restrict who has access to this material, which makes the privacy controls available through the social media website a concern.

A second concern is what a social media website can do with a person's data. It's not uncommon for websites to claim rights to aggregate, share, and sell users' personal data and content. While many issues surround this topic, this chapter will focus specifically on the privacy issues that arise from users' personal data being distributed by a website.

Related to both of these issues is the persistence of information. Once data becomes publicly available on the web, it is archived and cached by many different sites. Thus if users change their minds about what they want to share, previously posted information cannot effectively be removed from the web. This also applies to violations of users' privacy where information is shared without their consent or knowledge.

Problems facing social media users are increasingly arising from the information they share online. In some specific cases reported in the media, somepeople are seen to have suffered personal and professional problems related to their social media posts. A 24-year-old teacher was fired for posting a picture of

herselfon Facebook holding a glass of wine.[1] A teen was fired after she complained on Facebook that her job was "boring."[2] A Canadian woman had been diagnosed with major depression and was on disability for it. Her insurance company revoked her benefits, asserting that pictures on her Facebook page that showed her "having fun" were evidence that she was no longer depressed.[3]

More generally, social media content is growing as a source in legal proceedings. In 2009, a survey showed that Facebook was cited in 20% of U.S. divorce cases; in 2011, this increased to 33%.[4]

Employers also use social media to screen applicants. They have discovered that the information people share online is an easy alternative to a background check, and provides many insights into a person's character and activities. Screening via social media has become so popular that some companies have begun asking applicants—and some current employees—to provide their logins and passwords for Facebook and other websites. At the time of writing, this activity is being challenged in court, and the state of Maryland has passed a law making such inquiries illegal.

These trends indicate that controlling the privacy of the personal information we share on social networks is becoming increasingly important. Interest in that information and the associated risks are growing.

Understanding privacy first requires understanding the policies of websites, the technology of privacy, and best practices for sharing information online. The next sections will address these topics.

Privacy policies and settings

Social media, and social networks in particular, make it much easier for people to share information online. Before these social media technologies became widely available, some people created personal web pages, but the pages were harder to find, and the technical barrier of entry was relatively high for most people. Social media solved these problems by removing these barriers and providing a centralizing location to share and find information.

Privacy settings

People's comfort with sharing online (and the degree to which information is shared with a large audience) has changed over the course of the social media era. Consider as an example the evolution of Facebook's default privacy settings.

[1] http://www.ajc.com/news/barrow-teacher-fired-over-733625.html
[2] http://www.dailymail.co.uk/news/article-1155971/Teenage-office-worker-sacked-moaning-Facebook-totally-boring-job.html#ixzz1aP7zuSQ7
[3] http://www.dailyfinance.com/2009/11/23/facebook-spying-costs-canadian-woman-her-health-benefits/
[4] http://blog.divorce-online.co.uk/?p=2338

The site launched in 2004 and was originally restricted to people at universities. It opened to the general public in 2006.

In the early days of Facebook, most of a user's shared information was, by default, visible only to the user's friends. Over the course of its life, Facebook's default settings have changed to become increasingly public.

A visualization of the evolution of these default privacy settings is shown in the box later in this chapter.. This example explains how visible a user's data is with others online. This visibility is determined by the default privacy settings. As the figures show, the default settings on Facebook have become extremely public over time, with the defaults in mid-2012 allowing everyone on the Internet to see everything a person posts—with the exception of their contact information and their birthday.

This is not to say that a person cannot have a more private profile on Facebook. There are many privacy settings that can be used to restrict who sees what. More users have begun using these settings, too. A Pew study (Madden, 2012) showed that 58% of people have restricted access to their social networking profiles in some way. Unfortunately, the same study reports that many people have difficulty using and understanding these settings. Nearly half of the people interviewed reported having some difficulty managing the privacy controls.

The number and complexity of privacy management features varies widely between social media sites. Some have no options for privacy. This is especially common on review sites and social bookmarking sites. Large social networking sites (like Facebook and Google +), on the other hand, have many sophisticated tools for controlling privacy, sometimes allowing people to specify lists of individuals who have permission to see each individual piece of information. In between these extremes are sites that offer some limited controls. Twitter, for example, allows users to make their profiles public and visible to everyone, or private and visible only to approved followers.

Table 16.1 shows a matrix of some social media sites and the privacy settings they offer to users.

Privacy policies

But settings are just one piece of the privacy puzzle; in general, they affect only what other users are able to see on a person's profile. The information collected by sites, how it is used, and how it can be shared with other companies, is rarely controlled through privacy settings. Instead, this is detailed in privacy policies.

Of main concern with regard to privacy policies are the following issues:

What data is collected from users?

In order to establish an account, most websites require an email address and name. Some sites also collect location, photo, birthday, and other data. User's posts are also included here, since it is a type of personal data, but all

Table 16.1 Privacy Attributes of Various Social Media Sites

	Social Networking	Mictoblogging	Social Bookmarking	Photo Sharing	Cross-cutting	Location-Based Games	Marketplaces
	Facebook	**Twitter**	**Pinterest**	**Flickr**	**Google**	**FourSquare**	**Craigslist**
How is information collected							
From user	X	X	X	X	X	X	
From other websites (e.g. Facebook, twitter)			X			X	
Information shared by others about you	X	X			X		
Behavioral information (from logs, etc)	X	X	X	X	X	X	
What personal information is collected							
name	X	X	X	X	X	X	
email	X	X	X	X	X	X	X
location	X		X	X	X	X	
photo	X		X	X	X	X	
birthday	X			X	X	X	
posts (updates/text/ photos/etc)	X	X	X	X	X	X	X
How is information used							
registration	X	X	X	X	X	X	X
send email from the registering site	X	X	X	X	X	X	X
customer service	X		X	–	X	–	
recommendations (friends, products, etc)	X		X	–	X	–	

personalization	X		X	–	X	–		
sold				–		–		
Who is data shared with								
Other users on website - all			X				X	
Other users on the website - user controlled	X	X		X	X	X		
Other internet users (not registered with site)	X	X	X	X			X	
Third parties (other companies)	X	X	X	X	X	X		
For analysis provided back to registering site	X	X	X	X	X	X		
For marketing products to you	X			X		X		
For any purpose they choose	X					X		
Law enforcement if requested	–	X	X	X	X	X	X	
aggregated NPII data		X			X			
Companies that have an interest in the registering company	X	X	X	X	X	X		
User control issues								
Accounts/data can be totally deleted	X	X	X	X				
Archived copies kept			X		X		X	

sites do this since supporting user-generated content is at the core of the sites' functionality.

How is the data collected?

Sites will often collect data from users when they register. Others require users to link to other social networking accounts, like their Facebook and Twitter accounts. This makes it easier to share data on all platforms, and it also provides an additional source from which websites can harvest data about users.

Who is the data shared with?

The data users upload may be shared with other users and other companies. Privacy policies often stipulate if users have control over which specific people can see their data or if it is available to everyone on the site or everyone on the Internet. Policies will also detail which third parties can see the data. These may be companies who do analysis for the hosting website, marketing firms, or other sites that buy the data and use it for whatever purpose they like. The hosting company may give these third parties restricted access, sell their users' personal data, or give some of it away for free.

How is the data used?

Most sites use the personal information users provide to register them for the site and provide communication. They may also use it for customer service, personalizing the users' experiences, making recommendations, or supporting interaction.

What control does the user have?

If a user decides to delete his or her account, what rights do they have to how their data is handled? Some sites will delete all of the user's data and content. Others will keep archived copies for a fixed timeframe or in perpetuity. How account closing and data deletion is handled is usually addressed in the privacy policy.

Privacy policies are generally written in understandable plain English these days—an improvement from the times when they were full of legal jargon. Understanding what rights a social media site claims to personal information and content that users create should be an important factor in deciding what information to share.

Some sites have responded positively to actions that their users have undertaken in response to their privacy policies. For example, the social bookmarking tool Pinterest originally had a policy that claimed full ownership of any content that users uploaded, including the right to sell any of the images that were uploaded. This became a major issue for companies and professional photographers who wanted to retain rights to their images. According to these terms, even if someone else uploaded the photographer's image, Pinterest would claim rights to it. After a few months, Pinterest users began strongly objecting to these terms, and Pinterest removed the clause about ownership and the right to sell uploaded images.

Aggregation and data mining

Anonymous use of social media is possible, but remaining anonymous presents serious challenges. Many privacy policies speak of "personally identifiable information." This is data that reveals who you are, like your name or photo. Some data that is not useful on its own (like a ZIP code) may be combined with other data to become personally identifiable. Furthermore, a user may choose to share some information about herself, but not other information. Sophisticated techniques are being developed to allow third parties to infer some of the attributes that users have chosen not to share. This section provides an overview of some research being conducted in this arena.

Deanonymization

A small (but telling) study was conducted by Yates, Shute, and Rotman (2010). The researchers wanted to see how well protected the users' identities really were. This issue is relevant to many bloggers. They selected three anonymous bloggers. These people blogged under pseudonyms and tried to limit the information they shared about their families, places of work, and other personal details. A 2007 study (Qian and Scott, 2007) indicated that over 40% of bloggers censored their posts, including hiding their identities.

For their work, Yates et al. relied on the existence of marketing databases that will sell the name and address of everyone who meets the requestor's demographic requirements. For example, a requestor can specify a set of ZIP codes for location, age ranges, gender, marital status, and type of housing (rental, single family home, etc.). The database is intended for direct marketing and includes over 200 million Americans, with data compiled from a wide range of sources.

Is it possible to discover enough information about anonymous users that their identities could be discovered using a marketing database? The researchers read the blogs looking for the demographic information listed above. Gender was often easy to identify, as was marital status. For some users, a single ZIP code was easy to find because the blogger lived in a less populated area. For others, a set of ZIP codes for the blogger's home city were used. The blog posts also revealed what types of home each person lived in. Because the bloggers often posted about their birthdays, the researchers also found dates of birth for each person.

With this information in hand, the researchers queried the marketing database. Selecting everyone in the ZIP code range who matched the bloggers' age, gender, dwelling type, and marital status, they found that they could uniquely identify each person via their birthday with over 90% accuracy.

The research shows that online anonymity is very hard to maintain because only a few pieces of information—which appear to be meaningless for personally identifying someone—can be combined to reveal a person's identity.

If someone is using both anonymous and nonanonymous accounts (e.g., a professional account and an anonymous personal account), more sophisticated

computing techniques exist that can detect this and merge the two identities. These "entity resolution" computer algorithms use a combination of attributes, like addresses or birthdates, structural network data, and other features to merge nodes that represent the same person.

Inferring data

The approaches in the previous section are able to identify people who are anonymous or using multiple accounts. There are other techniques that use data people share in social media to infer more information about them.

One of the first such projects to receive wide media coverage was called gaydar. Developed as a term project at MIT, the application uses Facebook users' friend lists (publicly available by default) to predict the user's sexual orientation. In preliminary experiments, it was able to identify all the known homosexual men in their sample, even though these men had not listed their sexual orientation in their profiles. A similar tool, produced by Stockholm Pride, claims to analyze a person's Twitter posts and provide a "how hetero" score.[5]

Other researchers have used Twitter "following" relationships to identify people's political leanings. Golbeck and Hansen (2010) found the members of the U.S. Congress that a person followed, obtained a score of how liberal or conservative the congressperson was, and combined the scores of the congress people to come up with a score for the Twitter user. Combining users' scores to rate the political preferences of audiences for different media outlets produced results that closely matched previous studies of the media outlet's political leanings. Simple use of public following patterns yielded interesting insights into a user's politics.

Research has also shown that users' personality traits can be predicted with relative accuracy by using public profile data from Facebook and tweets posted on Twitter (Golbeck et al., 2011).

The above projects address data about individual users, but research has also shown that information about relationships is predictable. Gilbert and Karahalios (2009) used Facebook profile data and communication patterns to accurately predict tie strength between Facebook friends. Many researchers (Golbeck, 2009, Dubois et al., 2011, Ziegler, 2006) have shown that trust relationships can also be computed with some accuracy.

Data mining

The studies described so far infer information about specific traits. *Data mining*, on the other hand, uses many sophisticated computing techniques to discover previously unknown patterns and relationships in large collections of data. Data mining is used in many applications outside of social media as well. For example, one store used data mining on their sales receipts and found that men tended to buy diapers and beer together on Thursdays. Further, they found those families tended to do their main grocery shopping on Saturdays, so the Thursday trips were usually to

[5] www.stockholmpride.org/howhetero/

stock up on things for the weekend. This allowed the store to place a beer display closer to the diapers and ensure that they charged full price on Thursdays.

With social network data, companies will be looking for similar patters. Users with certain attributes may perform certain actions together. This can be used to target advertising to users, or to collect data on those users and sell it to third-party companies (so that *they* can directly market to the users).

Companies are already creating plans to mine social network data and use it in ways that people might not expect. In 2012, Germany's largest credit rating agency—which rates how likely people are to repay their loans and thus dictates the interest rates a person might receive on a credit card or their ability to get a mortgage—leaked news that it planned to use data from social media to identify potential customers and measure how risky they might be.[6] A public outcry about privacy issues shut the project down, but it indicates how information is available through social media.

Recommender systems also use data that people provide to make new suggestions. A person's ratings, reviews, and buying habits are all useful in making suggestions about new items that a user might like. Some recommender systems also use social data to improve these recommendations. Overall, research shows that users appreciate recommender systems; this example illustrates that technologies that use a person's information need not be threatening or scary.

Data ownership and maintaining privacy online

The interest of companies and organizations in users' data, the trend of social media toward making such information public by default, and the growing number of tools allowing others to discover new information can be overwhelming for social media users. Furthermore, even with well-tuned privacy settings, information shared online can almost never be considered truly private. Many sites have ways for clever or determined people to circumvent the privacy settings. Old data that a user may have deleted may still be archived on other sites. And perhaps the biggest (and technologically simple) threat is the following: Users with permission to see personal information can always copy it and share it with the wrong people.

A user can employ personal strategies to help keep social media data private. However, it is first important to know who owns the data shared through social media. Some websites allow users to own all the data they post. Flickr, the photo-sharing website, allows posters to retain ownership of everything they share. It also offers options for licenses, so that a user can dictate how others may use their photos. Other websites, like Wikipedia, require authors to give up ownership of their content as soon as it is posted on the site. Facebook technically allows users to maintain ownership of their data, but their terms of service state that you grant them "a non-exclusive, transferable, sub-licensable, royalty-free, worldwide

[6] http://moneyland.time.com/2012/06/14/could-that-facebook-like-hurt-your-credit-score/

license to use any IP content that you post." That means Facebook is allowed to do anything it wants with the data you upload, including selling it to other people, without paying you anything or asking for your consent.

The Facebook model is common among many social media websites. On one hand, it is important to these companies' business models that they can use people's data. Because most of these sites are free, they need to make money from someone other than the users. Most often, this comes from advertising, particularly from offering advertisers the opportunity to target very specific demographics, based on all the data users upload. In effect, social media users are not the customers of the social media companies; they are the product.

While these business models mean it is unlikely that social media will leave full control of personal data in the hands of users, it does not necessarily mean that the only solution is to stay offline. Understanding the privacy landscape allows users to make better decisions about what (and what not) to share. For example, privacy concerns are rarely voiced around the professional social network LinkedIn. That's largely because the information people put there is not sensitive; it is created for a professional audience, and it is intended to be seen by anyone on the Internet. Users make careful choices about what they post, and they know it will be public.

When using social media for personal rather than professional activities, people can still protect themselves. By default, assume that anything you post could find its way to your boss, potential employers (including jobs you will apply for years from now), all friends, and people who do not like you. Consider the repercussions of the information reaching those people. Then decide which things you are comfortable with reaching a large audience and how much to trust friends to protect those things. Being fully informed about who can see the information, how it can be used, and what the website's privacy policy is allow users to make the best decisions about what to share. And remember: once content is shared, it can never be fully retracted.

Respecting privacy in social media analysis

The wealth of information that people upload to social media websites has not been valuable just to the companies that use it for advertising and selling products. It has also become extremely useful to researchers who want to understand social behavior, relationships, user preferences, and most of the other things discussed in this book. How can researchers conduct their work analyzing social media while still respecting the privacy of users?

An overarching principle is to respect how users expect their information to be handled. When gathering data to analyze, it should be collected while respecting the terms of service and policies of the social media website. When users register and share their data, they know whether or not third parties (like researchers) are allowed to *scrape* their data (i.e., download it from the site). If the site forbids

scraping, that does not mean that it cannot technically be done. However, because it violates the terms of service for the website, and thus the expectations of users, it is generally not an ethical way to conduct research.

Many of these websites provide mechanisms for legitimately accessing user's data, often through an *application programming interface* or *API*. An API is a way to write a program that accesses social media information through the website's official channels. This will necessarily work in compliance with the website's terms of service, so the data will be collected in compliance with what the users expect.

Even though users often share data publicly and a researcher accesses it properly, users do not expect to see that data appear elsewhere online. Even when they grant a website license to use their data, users expect to be able to control what appears, delete things, and change their mind about how it is shared. When a third-party collects data and then shares it elsewhere, that control is taken from the users. At the same time, it may be scientifically important to share information that was properly collected in the course of analysis. Several strategies can help with this problem. First is data anonymization. Instead of revealing personally identifiable information about a person, anonymizing data and sharing only the relevant scientific details is often a good strategy. Also consider showing only aggregated data (averages, distributions, and other statistics) rather than the raw data. However, as was discussed above, sophisticated techniques can often deanonymize data. Thus, it is important to consider what information to share and to limit it as much as possible to protect the identity of users.

When more sensitive data is being used, or when it is more difficult to truly anonymize, it may be prudent to ask users to explicitly consent to their data being used. Informed consent is a critical part of the IRB (Institutional Review Board) process that is required at most universities. The IRB is a group of people who review all experimental protocols that involve human subjects. This includes research into social media. They require precise details about how public data will be collected, how it will be shared, and how the users' identity will be protected. When subjects are asked to interact with researchers or provide consent to use of their data, the IRB also reviews the procedures for collecting that consent and informing subjects of their rights to participate (or not). Anyone conducting social media analysis within a university will be required to outline their experimental protocols and receive IRB approval. Many companies, particularly those doing federally funded work, will also require approval from an IRB. This may be one established within the company or one that exists in a partner institution.

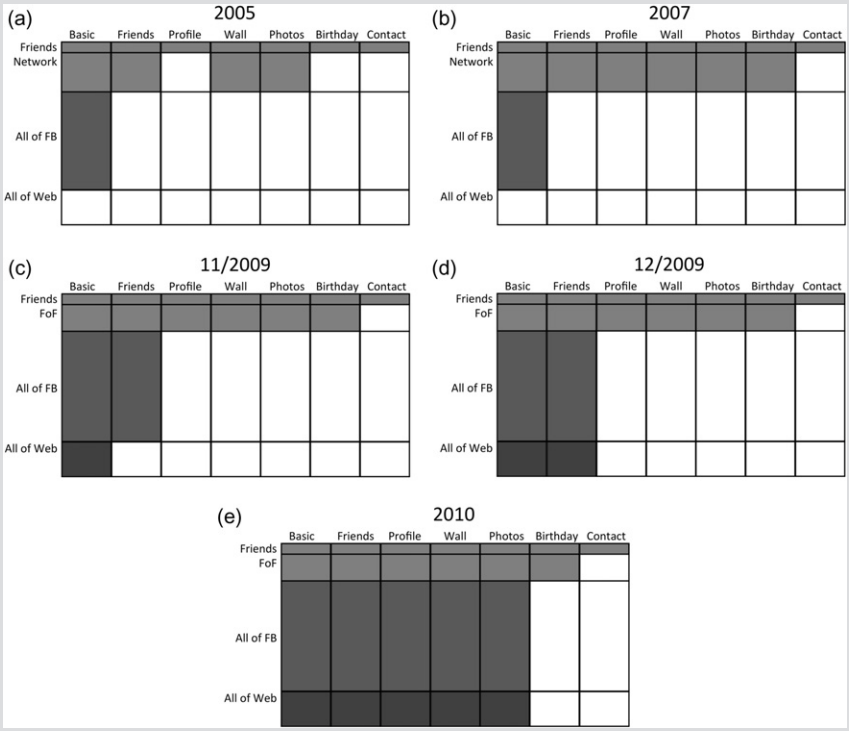
Following these guidelines will help researchers respect the privacy of users while still being able to conduct useful and interesting research. Particularly when research protocols go through IRB, researchers can be assured that their activities meet common ethical standards for protecting users' interests.

Exercises

1. Choose two social media websites. Pull up their privacy policies and answer the following questions for each.
   a. What information will be collected about you?
   b. How much of that information is really necessary for you to use the site? Are they asking for more than they need?
   c. Do you have access to all the information stored about you?
   d. How will your information be shared with third parties?
   e. Can your data be sold?
   f. Are you allowed to permanently delete your data from the system?
2. For the two systems you analyzed in exercise 1, which privacy policy makes you more comfortable and why?
3. Imagine you are running a social media website. Design your own privacy policy. Be sure it addresses the points listed in exercise 1.
   a. Is your policy geared more toward the user or toward your website?
   b. What things can you do with a person's data that may limit your company's opportunities?
   c. Conversely, are you granting more rights to the users than other sites do, and why?
4. Recall exercise 5 from Chapter 5 on tie strength. In that exercise, you looked at an online discussion group and listed ways to measure tie strength based on users' activity. Say you wanted to carry out that study and publish your results.
   a. What data would you want to collect?
   b. How would you collect it?
   c. Is your strategy in line with the discussion group's terms of service and privacy policy?
   d. Beyond what is available online in the discussion board, what information would you want to gather from users (e.g., through surveys or interviews)?
   e. What steps would you take to protect their privacy and prevent any of your results from being personally relatable back to the people?
5. Think about all the social media websites you belong to.
   a. If a company came along and bought your data from all of those companies and put it all together, what would your reaction be?
   b. What if that company decided to sell your aggregated data to marketers?
   c. What if they decided to sell it to the government?
   d. What if they decided to publish it freely online so that anyone could access it?
   e. Describe three threats you think people could face if their information were shared publicly?

EVOLUTION OF DEFAULT FACEBOOK PRIVACY SETTINGS

The amount of information shared publicly by default on Facebook changed dramatically between 2005 and 2010. This visualization illustrates the different types of data and their default settings over time.



Thanks to John Alexis Guerra Gómez and Cody Dunne for their help in producing these visualizations.

This page intentionally left blank

CHAPTER

17 Case Study: Social Network Strategies for Surviving the Zombie Apocalypse

Introduction

The zombies are coming

This theory, once relegated to the realm of speculative fiction, has now been accepted as an unfortunate eventual reality. Efforts to understand the phenomenon and prepare the population have been undertaken by scientists, the popular press, and government agencies such as the United States Centers for Disease Control and Prevention.[1]

To survive in a zombie-infested world, the individual survivor faces two major challenges:

- Avoid zombie infection.
- Maintain access to information.

On one level, these two goals are contradictory; the people providing helpful information may be the same ones to bite us in the end (quite literally).

In this chapter, we will apply lessons from social network analysis to develop life-preserving network-based strategies—including ideas about tie strength, network propagation, and network structure—for use by individuals to meet zombic challenges and survive to see another horror-filled, post-apocalyptic day.

Related work and background of the zombie apocalypse

Zombies have not been an important population in social networking, but there is an extensive body of work from other fields that describes the nature and traits of the undead. That literature is far too extensive to cover here, so we will review a summary of the major facts and theories required to support our work.

[1] http://www.cdc.gov/phpr/zombies.htm

We do not know when the zombie apocalypse will come, or what form the zombies themselves will take, but there are several schools of thought on the latter. All theories agree that zombies share several common traits: They display some level of anger or rage; they have an impulse to attack, harm, or eat the flesh of living humans; and they possess limited cognitive abilities, being generally unable to communicate. Because they do not react to pain, and often are already dead, they can be difficult to kill (or re-kill).

Zombies are evaluated according to three axes: alive/undead, speed, and intelligence. The major theories about how our zombie foes will manifest represent different blends of these three axes, the foremost of which are as follows:

- **Undead, slow, unintelligent zombies**
- In this scenario, zombies rise from the dead and are generally uncoordinated, ambling, unmotivated, and mostly unintelligent. They will pursue humans, but can be easily outrun and outsmarted. This theory of the coming horde was established in foundational work in the 1968 film *Night of the Living Dead*, and adopted by others (*Shaun of the Dead*, 2004, *Dawn of the Dead*, 2008).
- The zombies' numbers are the biggest threat with this group, as a large swarm can easily overpower an individual in spite of their physical limitations. Maintaining a strong core group of social connections is critical to survival in this case.
- **Undead, fast, somewhat intelligent zombies**
- Although zombies are never as intelligent as their human counterparts, in some scenarios they retain (or develop) some basic critical thinking skills. In the literature, this often correlates with increased speed and agility. Examples of this type of zombie can be found in *Day of the Dead* (2008), *Zombieland* (2009), and *Dead Snow* (2009).
- The landmark work, *The Zombie Survival Guide*, presents something of a hybrid view: Zombies respond to stimuli and seek out living victims, but display limited movement and agility. This is echoed, to some extent, in other work (e.g., *The Walking Dead*, 2011).
- When dealing with this type of zombie, communication among human groups is important, so that safe havens can be established and large zombie groups avoided.
- **Living, fast, virus-infected zombies**
- A theory brought forward most dramatically in 2002's 28 *Days Later*, this presents an alternative to the undead nature of zombies. This type of zombie occurs when a living human becomes infected with a virus that affects the brain, resulting in reduced thinking ability, increased rage, and often, increased speed and agility. Frankly, this zombie scenario is terrifying.
- With this type of zombie, all of our recommendations are of even greater importance, since the slightest contact with an infected zombie is extremely dangerous.

Note that for the purposes of this work, we will exclude zombies of voodoo origin (e.g., *The Serpent and the Rainbow*, 1988). Our concern here is with the spread of zombinism, which is not an issue among the excluded group.

For ease of narrative in this chapter, we will refer to zombinism as a "disease" and a person who is a zombie (or in the process of becoming one) as "infected," though this is not intended as an endorsement of any of the theories presented above.

In all the cases described above, zombinism spreads through an exchange of bodily fluids, often via a zombie bite of a human victim. In some cases, the change from human to zombie can be nearly instantaneous; at other times, the human may suffer a prolonged and painful death over many days before dying, and then be reanimated as a zombie. We choose to ignore the incubation period and treat zombinism as an *SI* disease model. This describes a disease where individuals in the population are susceptible to infection (*S*) and then become infected (*I*) without the opportunity for recovery and with no incubation period.

This chapter focuses on strategies an individual can use to improve his or her chances of survival, andpresents techniques the government can use to help as many citizens as possible.

Network strategies for the individual: Avoiding infection

When the zombie apocalypse begins, life will literally depend on avoiding infection. Fortunately, some important insights from social network literature are directly applicable in this context, and they have the potential to save lives.

Naturally, avoiding infection means avoiding zombies. While some advocate the lone wolf strategy, social connections can be extremely valuable in crisis times. At the same time, the more people with whom an individual comes into contact, the more chances there are for meeting zombies in their place. Fortunately, network analysis offers insights that can help guide survivors as to which connections they should eliminate and why.

Tie strength

The single most important concept from social network analysis we can use to improve chances of survival is tie strength (Granovetter, 1973).

Strong ties have been shown to be useful in maintaining emotional support (Schaefer et al., 1981) and in building a strong, supportive community. This can serve as an analog for the times of crisis that will certainly be upon us during the zombie apocalypse.

Strong ties are trusted and reliable connections. These are important in the time of the zombie apocalypse. However, weak ties have been shown to provide many social benefits. For any given person, their strong ties are likely to be connected to one another. Thus, the information that flows among strong ties tends

not to be novel since it was already known within the group. Weak ties, on the other hand, connect people to diverse groups of people who can provide new information and opportunities. Granovetter studied this topic in the context of job seeking, but it has since been shown that all types of things propagate through a network better along weak ties, including diseases (Gilbert, 2010).

Within the Human-Computer Interaction (HCI) community, social ties have also become an interesting area of study. Gilbert and Karahalios (2009) demonstrated how tie strength could be computed from social media websites. Relevant work was also done in Kivran-Swaine et al. (2011), who used tie strength to understand how users break relationships on Twitter by un-following one another—an action that will need to be enacted quite literally in the event of the zombie apocalypse. Similarly, Stutzman and Kramer-Duffield (2010) considered tie strength as a factor in privacy decisions on Facebook. Again, making determinations about privacy will be particularly important when zombies are present.

Extensive social interaction during the zombie apocalypse would put a person in contact with more weak ties, which in turn increases the likelihood of encountering someone who has (or will) become infected. Using these lessons, we find that a survivor must limit his or her physical contact with weak ties as much as possible. A close, trusted, and relatively small group, otherwise isolated from casual social contact, will be much less likely to encounter zombies and thus more likely to remain uninfected.

The zombie-filled world is unpredictable, however, and it is still possible that a strong tie may have an unfortunate chance encounter and become infected. In that case, the infection is likely to spread extremely quickly within the strong tie network. It is critical that the infected individual be killed as quickly as possible. We realize this is difficult advice to take, since the very nature of strong ties implies that the person will likely be a very close friend or relative. However, our study of the zombie canon suggests that the infected person will be better off dead than as a zombie, and the group survival depends on immediate elimination of the threat.[2]

Network structure

Reducing connections to weak ties, thereby reducing the number of overall connections in the network, is a critical strategy to be used in avoiding infection. However, there are idiots in the world, and not everyone has the strength of will to take the necessary actions to improve their chances for survival. Thus, there will remain some highly connected nodes in the social network. It is critical that a survivor eliminate these ties from their network, be they strong *or* weak, as the probability that these hubs will become zombies themselves is extremely high.

[2]*Counterexample:* John Leguiazamo's character in *Day of the Dead* who wanted to "see how the other half lives." He was a pretty awesome zombie, but not the kind of person you want in your strong tie network.

To illustrate the importance of both reducing one's number of contacts and cutting ties with network hubs, consider this simplified model of disease propagation. Let the probability that any node in the population is a zombie be p_z. Then, let the probability of becoming infected from a given infected neighbor be p_i. This probability should take into account a 100% transmission rate when bitten, but also the probability of avoiding contact. Then, for a node k with one neighbor, the probability of becoming infected is $p_z \times p_i$. Of course, in reality, the probability that any given node is infected will vary based on the node's behavior and environment, but the simplified assumptions will serve to illustrate this point.

A k-threshold model states that a disease must be transmitted by at least k neighbors for infection to occur. In the case of zombinism, clearly only one transmission point is required, so we use a 1-threshold model. We can compute the probability of infection either as the sum of all the scenarios where at least one neighbor is infected and also passes the disease, or as 1 minus the probability that none of the infected neighbors pass it. For simplicity of the formula, we will consider the latter.

For n neighbors, the probability of infection is given by:

$$1 - \sum_{j=0}^{n} p_z^{j*}(1-p_z)^{(n-j)*}(1-p_i)^j$$

For illustration, consider several sample probabilities. Let $p_z = 0.05$ and $p_i = 0.6$ (a scenario that would occur somewhat early in the outbreak with relatively quick-moving, somewhat intelligent zombies). Then for a node with five neighbors, the probability of infection is:

$$1 - \sum_{j=0}^{5} 0.05^{j*}(1-0.05)^{(5-j)*}(1-0.6) = 0.210$$

The probability of a node becoming infected with the same probabilities, but with 20 neighbors is 0.633, and with 100 neighbors it is 0.994. Clearly, hubs, which have many connections, increase their chance of becoming infected and thus of spreading zombinism to their neighbors.

These results show that *minimizing connections and cutting connections with* hubs is a critical strategy to reducing the chance of infection.

Network strategies for the government: Stopping the spread

In the early days of the outbreak, the government will have critical decisions to make if they want to stop the spread of zombinism and protect groups of people or areas of the country from being infected.

Their first important task will be to discover where the outbreak is happening. Are zombies spontaneously appearing around the world? Are the dead literally

rising from the grave? Or has there been a Patient Zero—a person who somehow "caught" zombinism before anyone else had it, perhaps in a mutation of a virus—and the disease is spreading out from that location?

Analyzing social media with location data can help track down where the zombies are and how fast they are spreading. Infrastructure for mobile and Internet communication will likely stay intact in the early stages of the outbreak, and people will most certainly be sharing their observations in social media.

Recall tools like Ushahidi where volunteers manually place reports of events on maps. The government could automate some of this process. For example, they could analyze Twitter posts for words like zombie, undead, and NOOOOOOOOOOO! Using either the user-provided location or tracking the location via the user's IP address, they could group the reports of zombic activity by county and visualize these on a map (see Figure 17.1). Smaller municipalities could potentially visualize down to a finer-grained locations, like neighborhoods or blocks, where the data is available.

These visualizations of social media data combined with location can be critical in deciding where to allocate resources. The map above clearly shows that the West Coast is unlikely to be saved. While California is a populous state, this map indicates that government resources would be better allocated to protecting states

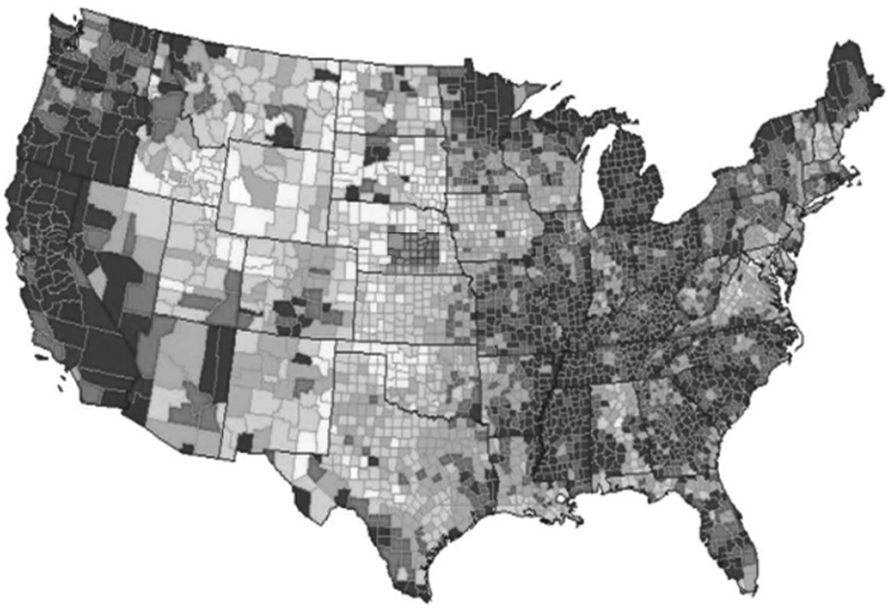


FIGURE 17.1

A possible map of zombie outbreaks. Dark colors indicate more reports of zombies. All nonzombies head to Wyoming!

from the Rocky Mountains out through the Great Plains and to closing the borders of at least the West Coast states, Minnesota, and Missouri.

Network strategies for the individual: Obtaining information

With reasonable precautions in place to avoid physical interaction with the people most likely to spread infection while maintaining a strong social group, the survivor's next concern should be access to information. Access to information can lead survivors to safe havens or additional resources; it can also warn of danger zones or degrading security situations. As discussed above, weak ties are usually the best source of this novel information—but how does one maintain that contact when, to ensure noninfection, one avoids contact with those same ties?

If the communication infrastructure remains in place, we obviously advocate extensive use of any and all social media channels. The effective use of social media in disaster scenarios—admittedly much smaller and simpler disasters than what we would face when the undead overrun the earth—has been demonstrated in many contexts, including Gunawan (2008), Palen and Liu (2007), and Starbird and Palen (2011). In particular, the discussion in Starbird and Palen's work (2011) of how Twitter users, via very short messages, were able to coordinate real and significant responsive actions to the 2010 Haiti earthquake is relevant. These lessons would directly translate into a zombie context.

Technology-based communications, the Internet, and eventually power grids are all likely to fail eventually, at which point traditional social media communication will not be available. In those cases, we must use technologies that are easy to build, maintain, and work in environments with limited power and communication infrastructure. The important network-based lesson here is that weak ties provide good sources of new information, and communicating with them will be important for organization, sharing resources, and receiving updates. When social media communications fail, it may still be possible to have communication over distance to reap the benefits of weak ties without risking physical contact.

Educational initiatives can train children in the zombie-free area to communicate without these technologies. Morse code is easy to learn and can be transmitted on radios that anyone can make with household items. Even if Morse code is only used for sending short messages, lessons from Twitter have shown that short messages can be quite effective in these types of scenarios. Second, we believe classes and experience in training carrier pigeons may be a prudent measure, to ensure our young people are capable of safe, long-distance communication.

Low-tech solutions can also be put in place to support community-maintained resources that will help survivors share information. For example, imagine geocaching where caches store information and other items useful to survivors. Instead of relying on GPS devices to find them, location information can be posted in central places (e.g., painted on billboards) or shared using the

communication mechanisms discussed above. Community members who rely on the caches will have incentives to maintain them; they can add their own information, rebuilding the cache when it is damaged, protecting it from weather, and informing others when and to where the cache has been moved due to increased zombie activity in the area.

Finally, trust must also be considered. Incorrect information can have dire consequences in a zombie-filled world. Bands of survivors should keep logs of the people who provide information and how trustworthy their information is. This generates a trust score for those sources. When social network data is available, it should also be recorded in these lots. Then, if information comes in from a new source, recipients can see if that new source is connected to any trusted people they know, and those trusted people can potentially vouch for (or discredit) the new person.

Network strategies for the government: Information sharing

The government will also have a great responsibility to share information with the public as quickly as possible. As the outbreak begins, most people will likely be tuning in to media of all types to receive information. As the zombies begin taking over and communications infrastructure becomes more limited, communication will likely have to be more limited. A good social network strategy is to dedicate resources to sharing with a small set of individuals who have the greatest reach within the social network.

For example, consider the network in Figure 17.2 that represents a small community in a zombie-safe area. If it is very expensive to establish communication with people, who are the few people with whom the government should invest?

First, consider the small groups around the edges of the visualization that are disconnected from the main component. These nodes will be unable to receive information shared with the large group. However, because resources targeting them would deprive the larger group, they are likely to be left isolated.

The question then becomes who within the giant component should receive and spread the information. Certainly, the government will want to target trustworthy, reliable individuals. However, without any information about individuals' reputations, the network structure can inform their decision. Nodes with high centrality by various measures will be important. Betweenness will identify people who are likely to transfer information between groups. Degree centrality will identify people who can directly reach many other people. Closeness centrality will identify people whose information will quickly reach a large number of people. Figure 17.3 shows individuals who have high centrality by all these measures marked with large black nodes in the network.

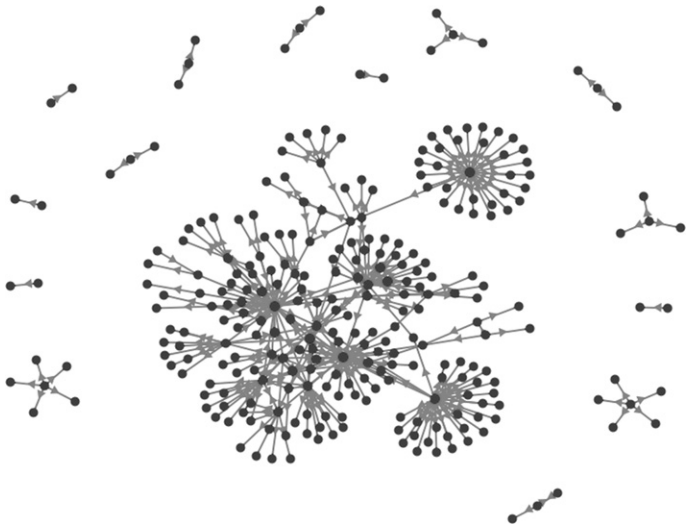


FIGURE 17.2

The social network of a community of zombie survivors.

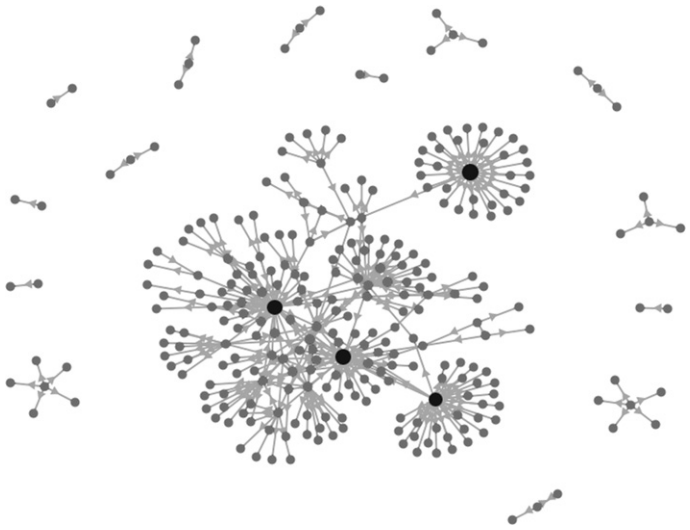


FIGURE 17.3

The four large black nodes have high centrality by several measures. Thus, they are ideal targets for the government to use to spread information.

These nodes are excellent choices for the government to use as information contacts. They have the contacts such that they can quickly and efficiently pass information to the majority of the giant component in only one or two steps. Even in larger networks, choosing people who have high centrality by many measures is the most effective way

Exercises

1. Aside from the geocaching example described above, think of another information-sharing community-maintained resource that a group of survivors could build during the zombie apocalypse. How would they share the information using only low-tech solutions? What is the motivation for people to participate in maintaining the resource?
2. In addition to the county-by-county map of the zombie outbreak presented in Figure 17.1, describe a visualization the government could use to track the spread of zombinism. Create an example of visualization with fake data (you could make the data up or use data about another disease outbreak). Explain the visualization and what it shows.
3. Imagine the zombies arrive tomorrow. Describe your survival strategy and list three social network insights that relate to your efforts.
4. Imagine you are an Evil Warlord in the post-apocalyptic zombie era. You have captured a zombie and you plan to sneak it into the community run by your nemesis and release it. You have a highly accurate adjacency list of the social network connections in that community.
   a. Into whose house should you release the zombie if your goal is to infect as many people as possible? Assume the person you infect bites many of his or her social connections, turning them into zombies, they do the same, and so on. Describe the network statistics you would compute.
   b. Download the Nemesis Network dataset from the book website and open it in your favorite network analysis tool (e.g., NodeXL or Gephi).
   c. Create a visualization of the network.
   d. Compute the statistics you identified in part a on this network.
   e. Give the ID number of the person whom you would attack with the zombie.
   f. Explain why you made these choices and why you think the zombinism would reach many people this way.
5. Imagine you are in charge of the community being infected and you have learned the Evil Warlord's strategies from exercise 4. You were unable to stop the zombie from infecting the target, but now you want to stop the spread. Imagine the infected node can bite two people at a given time step. After that, you can isolate three people in the community. Then the zombies each bite two people again.

With a partner playing the Evil Warlord and you playing the Nemesis, print out the visualization of the network. Let the Evil Warlord infect someone. That person will bite two neighbors (chosen by the Evil Warlord in this simulation), and then you can protect three people. Repeat this process until the spread stops. How many people have you saved, and how many people are infected?

This page intentionally left blank

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, *25* (3), 211–230.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2011). Four degrees of separation. *Computing Research Repository (CoRR)*, *abs/1111.4*.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Blau, J. R. (1995). When weak ties are structured. *Social Roles And Social Institutions: Essays in honor of Rose Laub Coser*, 133–147.

Brizan, D. G., & Tansel, A. U. (2006). A survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, *6*(3), 41–50.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564. doi:10.1086/630200(http://cid.oxfordjournals.org/content/49/10/1557.abstract)

Cook, K. (Ed.), (2001). *Trust in Society,* New York: Russell Sage Foundation.

Deutsch, M. (1962). Cooperation and trust. Some theoretical notes. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation*. Nebraska University Press.

Dunne, C., & Shneiderman, B. (2009). (HCIL Tech Report HCIL-2009-13) *Improving graph drawing readability by incorporating readability metrics: A software tool for network analysts.* University of Maryland.

Dunne, C., & Shneiderman, B. (November 2012). Motif simplification: Improving network visualization readability with fan and parallel glyphs. HCIL Tech Report.

Eleta, I. (2012). Multilingual use of twitter: Social networks and language choice, *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*. ACM.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*(6), 1585–1593.

Forte, A., & Bruckman, A. (2005). Why do people write for wikipedia? Incentives to contribute to open-content publishing. *SIGGROUP 2005 Workshop: Sustaining Community*.

Freeman, L.C. (1979). Centrality in social networks conceptual clarification. *Social Networks,* 1(3), 215–239. doi:10.1016/0378-8733(78)90021-7. <http://www.sciencedirect.com/science/article/pii/0378873378900217>.

Gilbert, E.E. (2010). Computing tie strength. PhD Dissertation, University of Illinois.

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the 27th international conference on human factors in computing systems*, 211–220. New York, NY, USA: ACM. doi:10.1145/1518701.1518736. <http://doi.acm.org/10.1145/1518701.1518736>.

Glott, R., Schmidt, P., & Ghosh, R. (2010). *Wikipedia survey—overview of results. United Nations University: Collaborative Creativity Group*.

Goel, S., Muhamad, R., & Watts, D. J. (2009). Social search in 'Small-World' experiments. In *Proceedings of the 18th international conference on world wide web*, 701–710. New York, NY, USA: ACM. doi:10.1145/1526709.1526804. <http://doi.acm.org/10.1145/1526709.1526804>.

Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Trans. Web* 3(4):12:1−12:33. doi:10.1145/1594173.1594174. <http://doi.acm.org/10.1145/1594173.1594174>.

Golbeck, J., Grimes, J. M., & Rogers, A. (2010). Twitter use by the U.S. Congress. *Journal of the American Society for Information Science and Technology*, *61*(8), 1612−1621.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K., (2011). Predicting Personality from Twitter. *In Proceedings of the 3rd IEEE International Conference on Social Computing*, (pp. 149−156). Boston, Massachusetts, USA.

Golembiewski, R., & McConkie, M. (1975). The centrality of interpersonal trust in group processes. In C. Cooper (Ed.), *Theories of Group Processes*. Hoboken, NJ: Wiley.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78* (6), 1360−1380.

Gunawan, L. T. (2008). Collaboration-oriented design of disaster response system. In *CHI '08 extended abstracts on human factors in computing systems*, 2613−2616. New York, NY, USA: ACM. doi:10.1145/1358628.1358727. <http://doi.acm.org/10.1145/1358628.1358727>.

Hardin, R. (2002). *Trust & Trustworthiness,* New York: Russell Sage Foundation.

Harel, D., & Koren, Y. (2000). A fast multi-scale method for drawing large graphs. In Joe Marks (Ed.), *Proceedings of the eighth international symposium on graph drawing (GD '00)*. London, UK: Springer-Verlag (183−196).

Johnson-George, C., & Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, *43*(6), 1306.

Karau, S. J., & Williams, K. D. (2001). Understanding individual motivation in groups: The collective effort model. *Groups at Work: Theory and Research*, 113−141.

Karweit, N. L. (1979). The Conditions for Peer Associations in School. Report 282, Center for Social Organization in Schools, Johns Hopkins University.

Kivran-Swaine, F., Govindan, P., & Naaman, M. (2011). The impact of network structure on breaking ties in online social networks: unfollowing on twitter. In *Proceedings of the 2011 annual conference on human factors in computing systems*, 1101−1104. New York, NY, USA: ACM. doi:10.1145/1978942.1979105. <http://doi.acm.org/10.1145/1978942.1979105>.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 72.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Trans. Web* 1(1). doi:10.1145/1232722.1232727. <http://doi.acm.org/10.1145/1232722.1232727>.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining*, 631−636. New York, NY, USA: ACM. doi:10.1145/1150402.1150479. <http://doi.acm.org/10.1145/1150402.1150479>.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019−1031.

Lin, N., Dayton, P. W., & Greenwald, P. (1978). Analyzing the instrumental use of relations in the context of social structure. *Sociological Methods & Research*, *7*(2), 149−166.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). I'm the mayor of my house: Examining why people use foursquare—a social-driven location sharing application. In *Proceedings of the 2011 annual conference on human factors in computing systems*, 2409−2418. New York, NY, USA: ACM, 2011. doi:10.1145/1978942.1979295. <http://doi.acm.org/10.1145/1978942.1979295>.

Livne, A., Simmons, M. P., Adar, E., & Adamic, L. A. (2010). *The party is over here: Structure and content in the 2010 election* (201−208) *Proceedings of the fifth international AAAI conference on weblogs and social media.* MI, USA: Ann Arbor.

Lomnitz, L. (1977). *Networks and Marginality,* New York: Academic Press.

Longueville, B. D., Smith, R. S., & Luraschi, G. (2009). 'OMG, from here, i can see the flames!': A use case of mining location based social networks to acquire spatiotemporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks*, 73−80. New York, NY, USA: ACM. doi:10.1145/1629890.1629907. <http://doi.acm.org/10.1145/1629890.1629907>.

Madden, M. (2012). Privacy management on social media sites. *Pew Internet and American Life Project*.

Marsden, P. (1984). Measuring tie strength. *Social Forces*, *63*(2), 482−501.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, *23*(3), 473−490.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2000). Trust in e-Commerce vendors: A two-stage model. In *Proceedings of the 21st International Conference on Information Systems*, 532−536. Atlanta, GA, USA: Association for Information Systems. <http://dl.acm.org/citation.cfm?id=359640.359807>.

Metaxas, P.T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (Not) to predict elections. In: *Proceedings of IEEE SocialCom/PASSAT 2011* (pp. 165−171). <http://ieeexplore.ieee.org/xpl/login.jsp?tp = &arnumber = 6113109&url = http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6113109>.

Milgram, S. (1967). The small world problem. *Psychology Today*, *2*(1), 60−67.

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P.T. (2011). Vocal minority versus silent majority: Discovering the opinions of the long tail. In: *2011 IEEE third international conference on social computing (2011)* (pp. 103−110). <http://cs.wellesley.edu/∼pmetaxas/Silent-minority-Vocal-majority.pdf>.

Neustaedter, C., Tang, A., & Tejinder, J. K. (2010). The role of community and groupware in geocache creation and maintenance. In *Proceedings of the 28th international conference on human factors in computing systems*, 1757−1766. New York, NY, USA: ACM. doi:10.1145/1753326.1753590. <http://doi.acm.org/10.1145/1753326.1753590>.

Nov, O. (2007). What motivates wikipedians? *Communications of the ACM* 50(11):60−64. doi:10.1145/1297797.1297798. <http://doi.acm.org/10.1145/1297797.1297798>.

O−Hara, K. (2008). Understanding geocaching practices and motivations. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, 1177−1186. New York, NY, USA: ACM, 2008. doi:10.1145/1357054.1357239. <http://doi.acm.org/10.1145/1357054.1357239>.

Palen, L., & Liu, S. B. (2007). Citizen communications in crisis: Anticipating a future of ICT-supported public participation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 727−736. New York, NY, USA: ACM. doi:10.1145/1240624.1240736. <http://doi.acm.org/10.1145/1240624.1240736>.

Plaisant, C., Grosjean, J., & Bederson, B.B. (2002). SpaceTree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In: *Procedings of IEEE symposium on information visualization, 2002* (pp. 57−64). Boston, MA.

Qian, H., & Scott, C. R. (2007). Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication*, *12*(4), 1428−1451.

Ratkiewicz, J., Conover, M., Meiss, M., Gonc¸alves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, 249−252. New York, NY, USA: ACM. doi:10.1145/1963192.1963301. <http://doi.acm.org/10.1145/1963192.1963301>.

Schaefer, C., Coyne, J. C., & Lazarus, R. S. (1981). The health-related functions of social support. *Journal of Behavioral Medicine*, *4*(4), 381−406.

Stack, C. B. (1975). *All our kin: Strategies for survival in a black community,* Basic Books.

Starbird, K., & Palen, L. (2011). 'Voluntweeters': Self-organizing by digital volunteers in times of crisis. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, 1071–1080. New York, NY, USA: ACM. doi:10.1145/1978942.1979102. <http://doi.acm.org/10.1145/1978942.1979102>.

Strange, H. (2011). *Facebook and location services,* University of Copenhagen.

Stutzman, F., & Kramer-Duffield, J. (2010). Friends only: Examining a privacy-enhancing behavior in facebook. In *Proceedings of the 28th international conference on human factors in computing systems*, 1553−1562. New York, NY, USA: ACM. doi:10.1145/1753326.1753559. <http://doi.acm.org/10.1145/1753326.1753559>.

Sztompka, P. (1999). *Trust: A sociological theory,* Cambridge University Press.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In: *Proceedings of the fourth international AAAI conference on weblogs and social media* (pp. 178−185).

Turner, T. C., Smith, M. A., Fisher, D., & Welser, H. T. (2005). Picturing usenet: mapping computer-mediated collective action. *Journal of computer-mediated communication* 10(4):0. doi:10.1111/j.1083-6101.2005.tb00270.x. <http://dx.doi.org/10.1111/j.1083-6101.2005.tb00270.x>.

Valdivia, A., et al. (2010). Monitoring Influenza Activity in Europe with Google Flu Trends: Comparison with the Findings of Sentinel Physician Networks-Results for 2009−10, 15(29), 2.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'Small-world' networks. *Nature*, *393*(6684), 440−442.

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. A. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, *8*(2), 1−32 (http://www.cmu.edu/joss/content/articles/volume8/Welser/).

Wu, P., Qu, Y., Preece, J., Fleischmann, K.R., Golbeck, J., Jaeger, P.T., & Shneiderman, B. (2008). Community response grid (CRG) for a university campus: Design requirements and implications. In: *Proceedings of the fifth international conference on information systems for crisis response and management (ISCRM'08).* Washington, DC, USA.

Yang, H. -L., & Lai, C. -Y. (2010). Motivations of wikipedia content contributors. *Computers in human behavior* 26(6):1377−1383. doi:10.1016/j.chb.2010.04.011. <http://www.sciencedirect.com/science/article/pii/S0747563210000877>.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260−281.

Yates, D. N., Shute, M., & Rotman, D. (2010). Connecting the dots: When personal information becomes personally identifying on the internet. *International AAAI Conference on Weblogs and Social Media (ICWSM'10)*.

Ziegler, C. -N., & Golbeck, J. (2007). Investigating interactions of trust and interest similarity. *Decision support systems* 43(2):460−475. doi:10.1016/j.dss.2006.11.003. <http://www.sciencedirect.com/science/article/pii/S0167923606001655>.

This page intentionally left blank

Glossary

Absent tie A nonrelationship between two people. This may describe people who do not know one another at all, or people who may pass on the street and wave but who do not otherwise have a relationship.

Adjacency list A list of node pairs that have a relationship, in the form of a,b" where a and "b are names of nodes.

Adjacency matrix A table listing all nodes on both the x- and y-axis and indicating the presence of an edge between two nodes in the cell where their row and column intersect.

Affiliation network A network that connects people to organizations.

Betweenness A measure of centrality that indicates how often a node lies on the shortest paths between all other pairs of nodes in the network.

Bimodal graph A network with two types of nodes, for example, nodes representing people and organizations.

Bipartite graph A bimodal graph where nodes of one type are only connected to nodes of the other type.

Bridge An edge that, if removed, would increase the number of components in the network.

Centrality There are several different methods for computing centrality, but all are designed to indicate the importance of a node in the network.

Centralization A measure of how importance, measured by centrality, is distributed in the network.

Clique A set of nodes that are all connected to each other. The density of a clique is 1.

Closeness A centrality measure that uses the average shortest path length from the selected node to every other node in the network.

Cluster A grouping of nodes that are similar to one another or closely connected.

Clustering coefficient The density of a node's 1.5 diameter egocentric network.

Cohesion A measure of the minimum number of nodes that would need to be removed from the network before it becomes disconnected.

Community-maintained resource A website or service where content and administration are all handled by members of the community.

Compartmental models These models are used to study the spread of disease in a network. People are compartmentalized according to their disease state (e.g., infected, recovered).

Connected component A graph or subgraph where every pair of nodes is connected with a path.

Connectivity See Cohesion.

Crowdsourcing Outsourcing tasks to a group of people online.

Degree The number of edges connected to a node.

Degree centrality Centrality measured by the degree of the node.

Degree distribution A chart indicating the number of nodes for each degree in the network.

Density The ratio of the number of edges in a network to the number of possible edges in the network.

Directed network A network with directed edges, which indicate a relationship from one person to another that is not necessarily reciprocal.

Edge A connection between two nodes; also called a link.

Egocentric network For a given node, its egocentric network consists of its neighbors and the edges to them in the network.

Eigenvector centrality A centrality measure that scores nodes based on the principle that relationships with more important nodes confer more importance than relationships with less important nodes. It is computed by calculating an eigenvector on the adjacency matrix.

Emotional intensity A measure of the closeness of the relationship.

Emotional support Communication between people that validates their feelings.

Epidemic models A means of describing the spread of a disease through a network.

Erdős number On the network of co-authors of scientific publications, the Erdős number indicates the shortest path between an author and Paul Erdős, a famous graph theorist and mathematician.

Facebook The largest social network on the web with over 1 billion users

Firefighter problem A method for thinking about containing the spread of disease in a network, where nodes are treated as "trees" and a fire spreads from tree to tree across edges. "Firefighters" can block edges to prevent the spread of a fire.

Forbidden triad A group of three nodes where two pairs are connected with strong ties and the third pair of nodes has no connection at all. It is called "forbidden" because it is unlikely to naturally occur in social networks.

Force directed layout A mechanism for laying out a graph for visualization. Nodes and edges are treated like physical objects and forces, and a model of those forces is used to determine positioning.

Foursquare A location-based game and service where users can "check in." to various locations, earn points, and share their location with friends.

Frictionless sharing Coined by Facebook, "frictionless sharing" describes how information is shared socially without any active choice to share by the user.

Graph A collection of nodes and edges, also known as a network.

Heterogeneous graph A graph with more than one type of graph.

Hub A node with many edges.

In-degree The number of direct edges coming in to a node.

information visualization The study of visual representations of data to aid in the discovery of patterns and understanding of the information.

Intimacy Mutual confiding or secret sharing. This is an indicator of tie strength.

k-threshold model An epidemiological model that indicates how many neighbors of a node must be infected for the disease to be passed on. The number of infected nodes must be at least "k."

Label A description of an edge, often the type of relationship it represents.

Link A connection between two nodes; also called an edge.

Location-based resources Online resources that take advantage of users' location information, often obtained by GPS or cellular signals on a mobile decide.

Microblog A service that allows users to post very short messages and updates.

Multimodal network A network with more than one type of node. For example, the network may have nodes representing people, schools, and cities.

Multiplex network A network with multiple types of edges.

Network A collection of nodes and edges, also known as a graph.

Node The representation of an entity in a network, usually a person in a social network; also known as a vertex.

Out-degree In a directed graph, the number of edges going out from a node.

PageRank A core of Google's ranking algorithm, which uses eigenvector centrality over the network of web pages and links on the web.

Path A series of edges that connect two nodes. It may be a direct connection or a series of steps with intermediate nodes.

Random graph A graph where edges are added between nodes at random.

Reciprocal services Favors that people do for one another, an indicator of tie strength.

Recommender system A system that tries to predict items that a user will like based on their previously expressed preferences.

Return on investment A measure of profits relative to investment.

ROI See Return on Investment.

Shortest path length In a graph, the path with the fewest number of edges that connects two nodes.

SI An epidemic model where people are either susceptible to a disease (S) or infected (I).

Singleton A node with no edges.

SIR An epidemic model where people are either susceptible to a disease (S), infected (I), or recovered/removed (R).

SIRS An epidemic model where people are either susceptible to a disease (S), infected (I), or recovered/removed (R). After recovery, the node may become susceptible again.

SIS An epidemic model where people are either susceptible to a disease (S) or infected (I). After the infection passes, the node becomes susceptible again.

Six degrees of separation The notion that any two people are connected by a path with an average length of six.

Small world A network characterized by a short average shortest path length and a high average clustering coefficient.

Social distance The difference in people's social backgrounds with respect to factors like education, income, and race.

Social network A network where the nodes generally represent people and the edges represent their social relationships.

Strong tie A close, strong relationship.

Strongly connected component A component in a directed network where a path following the correct direction of the edges exists between all nodes.

Subnetwork A subset of nodes and edges in the network.

Tie strength The closeness of a relationship between people.

Time The amount of time people spend together, have known one another, and the frequency of their interactions. An indicator of tie strength.

Trust A belief in someone's good intentions and a willingness to make a commitment based on that belief.

Twitter The most popular microblogging service.

Undirected edge An edge that reflects a mutual relationship.

Undirected network A network with undirected edges.

Vertex The representation of an entity in a network, usually a person in a social network; also known as a node.

Weak tie A casual relationship, often an acquaintance.

Weakly connected component A component in a directed network where a path exists between all nodes when the edges are treated as undirected.

Weight A numerical value on an edge, often reflecting the intensity of the relationship or frequency of interaction.

Wiki Software that supports collaborative creation and editing of websites.

Zombie An undead person, usually with very limited cognitive abilities and an appetite for human flesh (brains and otherwise).

Index

Note: Page numbers followed by "*f*" and "*t*" refer to figures and tables, respectively.

E

F

G

H

I

O

P

R

S

T

U

V

W

X

Y

Z