CHAPTER

3

Social Network Analysis

Measuring, Mapping, and Modeling Collections of Connections

OUTLINE

3.1 INTRODUCTION

Human beings have been part of social networks since our earliest days. We are born and live in a world of connections. People connect with others through social networks formed by kinship, language, trade, exchange, conflict, citation, and collaboration. Computer technologies used to create social networks are relatively new, but networks of social interactions and exchanges are primordial. Simply stated, a network is a collection of things and their relationships to one another. The "things" that are connected are called nodes, vertices, entities, and in some contexts people. The connections between the vertices are called edges, ties, and links. Many natural and artificial systems form networks, which exist in systems from the atomic level to the planetary level. *Social* networks are created whenever people interact, directly or indirectly, with other people,

institutions, and artifacts. Social network theory and analysis is a relatively recent set of ideas and methods largely developed over the past 80 years. It builds on and uses concepts from the mathematics of graph theory, which has a longer history. Using network analysis, you can visualize complex sets of relationships as maps (i.e., graphs or sociograms) of connected symbols and calculate precise measures of the size, shape, and density of the network as a whole and the positions of each element within it.

The recent proliferation of Internet social media applications and mobile devices has made social connections more visible than ever before (Chapter 2). The idea of networks, whether they are composed of friends, ideas, or web pages, is increasingly an important way to think about the modern world. Social network analysis helps you explore and visualize patterns found within collections of linked entities that include people. From the perspective of social network analysis, the treelike "org-chart" that commonly represents the hierarchical structure of an organization or enterprise is too simple and lacks important information about the cross connections that exist between and across departments and divisions. In contrast with the simplified tree structure of an org-chart, a social network view of an organization or population leads to the creation of visualizations that resemble maps of highway systems, airline routes, or rail networks (See Chapter 8). Social network maps can similarly guide journeys through social landscapes and tell a story about how some points are at the center or periphery of the network. Transportation networks where distance is measured in number of flights or roads from one city to another city are familiar. They inspire application to less familiar networks of electrical connections, protein expression, and webs of information, conversation, and human connection.

Social network analysis and metrics are described in several excellent books and journals [1–6]. This chapter touches on the key historical developments, ideas, and concepts in social network analysis and applies them to social media network examples. We have left details of advanced topics and mathematical definitions of various concepts to the many fine technical works. The following is intended as an introductory survey of the core network concepts and methods used in subsequent chapters, which focus on the networks that can be extracted from social media sources like Twitter, Facebook, email, discussion forums, YouTube, Flickr, wikis, and the web.

3.2 THE NETWORK PERSPECTIVE

Network analysts see the world as a collection of interconnected pieces. Those studying social networks see *relationships* as the building blocks of the social world, each set of relationships combining to create emergent patterns of connections among people, groups, and things. The focus of social network analysis is between, not within people. Whereas traditional social science research methods such as surveys focus on individuals and their attributes (e.g., gender, age, income), network scientists focus on the connections that bind individuals together, not exclusively on their internal qualities or abilities. This change in focus from attribute data to relational data dramatically affects how data are collected, represented, and analyzed. Social network analysis complements methods that focus more narrowly on individuals, adding a critical dimension that captures the connective tissue of societies and other complex interdependencies.

Network analysis shares some core ideas with the real estate profession. In contrast to approaches that look at internal attributes of each individual, network analysis shares the real estate focus on location, location, location! The interior of a house may be a liability, but where a property is located matters far more when trying to get a good sale price. The network perspective looks at a collection of ties among a population and creates measurements that describe the location of each person or entity within the structure of all relationships in the network. The position or location of a person or vertex in relation to all the others is a primary concern of social network analysis. Many network explanations look for causes of outcomes in the patterns of connections around an individual instead of their personal characteristics. "Know who" is often more important in network explanations than "know how." Network approaches observe that different people in similar circumstances and social positions often act in similar ways. Positions within networks may be as significant a factor as any aspect of the people who occupy them. Network analysis argues that explanations about the success or failures of organizations are often to be found in the structure of relationships that limit and provide opportunities for interaction [7].

Many network concepts are intuitive and echo familiar phrases like "friend of a friend," "word of mouth," and "six degrees of separation." Other network terms like "transitivity," "triadic closure," and "centrality" (see Section 3.5) may be unfamiliar terms for familiar social arrangements. Many of us recognize social network differences among people: we know some people who are "popular" and have connections to many others. We may also know some people who may be less "popular" but are still "influential," connecting to a smaller number of people who have "better" connections. Network analysis recognizes these and other less intuitively sensed patterns in social relationships, like measuring the number of your friends who know each

other and how much a person occupies a gatekeeper or bridge role between two groups. The network analysis approach makes the web of interconnections that bind people to one another visible, creating a mathematical and graphical language that can highlight important people, events, and subgroups.

3.2.1 A Simple Twitter Network Example

To better understand the network perspective, consider the social network of Twitter users shown in Figure 3.1 (see Chapters 2 and 10 for a description of Twitter). It is an example of a sociogram, also called a network graph, which is a common way of visualizing networks. Like all networks, it consists of two primary building blocks: vertices (also called nodes or agents) and edges (also called ties or connections). The vertices are represented by images of the Twitter users, and the edges are represented by the lines that point from one vertex to another.

This simple graph paints a picture of the social relationship of the Twitter users who tweeted about a 2009 workshop on information in networks at New York University[1] by including the text string "#WIN09." The size of each Twitter user's profile image is determined by the user's total number of tweets as reported by the Twitter Application Programmer Interface (API), which gives sophisticated users access to powerful services. This is one example of how attribute data (e.g., data that describe a person) can be overlaid onto a network. A line, or edge, exists between two people when one "follows" the other or if one user "mentions" or "replies" to the other. All of these connections in aggregate reveal the emergent structure of two distinct groups with few connecting links. This accurately represents the way the workshop brought together previously separate clusters of people from different disciplines. It also helps identify individuals who fill important positions in the network, such as those who many people follow and those who are connected to both clusters. This and following

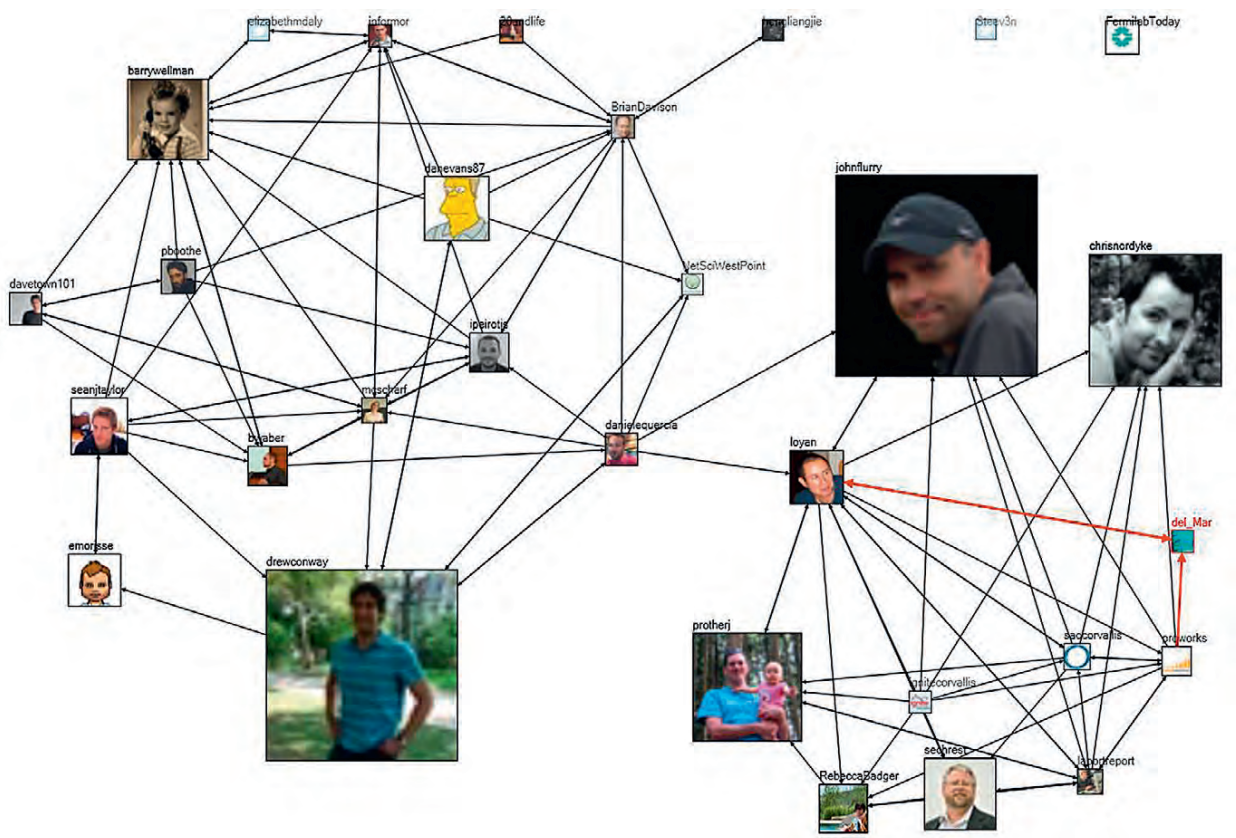

FIGURE 3.1 A NodeXL social media network diagram of relationships among Twitter users mentioning the hashtag "#WIN09" used by attendees of a conference on network science at New York University in September 2009. The size or each user's vertex is proportional to the number of tweets that user has ever made.

[1]http://winworkshop.net

chapters will provide a guide to creating maps like these from Twitter and other social media platforms and data sources. For now, let's consider the major components of a network in a bit more detail.

3.2.2 Vertices

Vertices, also called nodes, agents, entities, or items, can represent many things. Often they represent people or social structures such as workgroups, teams, organizations, institutions, states, or even countries. At other times they represent content such as web pages, keyword tags, or videos. They can even represent physical or virtual locations or events. They often correspond with the primary building blocks of social media platforms as described in Chapter 2: pages in wikis, friends in social networking sites, and posts or authors in blogs.

Although not necessary for network analysis, having attribute data that describe each of the vertices can add insights to the analysis and visualizations. For example, Figure 3.1 used descriptive attribute data about the total number of posts to convey a sense of who is most active on Twitter. Other attribute data from Twitter, such as the number of followers, people they follow, and their join date, can also be mapped to visual attributes (see Chapter 10). More generally, attribute data may describe demographic characteristics of a person (age, gender, race), data that describe the person's use of a system (number of logins, messages posted, edits made) or other characteristics such as income or location. In network visualization tools such as NodeXL, attribute data can be mapped to visual properties such as the size, color, or opacity of the vertices (see Chapter 4).

3.2.3 Edges

Edges, also known as links, ties, connections, and relationships, are the building blocks of networks. An edge connects two vertices together. Edges can represent many different types of relationships like proximity, collaborations, kinship, friendship, trade partnerships, citations, investments, hyperlinking, transactions, and shared attributes. A tie can be said to exist if it has some official status, is recognized by the participants, or is observed by exchange or interaction between them. A tie is any form of relationship or connection between two entities.

Network scientists have developed a language to describe different types of edges. In Section 2.3.5 of Chapter 2, we introduced the core types of connections that occur in social media networks. Here we describe how those concepts map to network and graph theory concepts more generally.

Undirected or *directed* edges are the two major types of connections. Directed edges (also known as asymmetric edges) have a clear origin and destination: money is lent from one person to another, a Twitter user follows another user, an email is sent to a recipient, or a web page links to another web page. They are represented on a graph as a line with an arrow pointing from the source vertex to the recipient vertex (see Figure 3.1). Directed edges may be reciprocated or not. If I sent you a message you may send one back in return, or not. An undirected edge (also known as a symmetric edge) simply exists between two people or things: a couple is married, two Facebook users are friends, or two people are members of the same organization. No origin or destination is clear in these mutual relationships. They cannot exist unless they are reciprocated. Undirected edges are represented on a graph as a line connecting two vertices with no arrows.

Edges can be represented by different types of data. The simplest type of edge, an *unweighted* edge or binary edge, only indicates if an edge exists or not. For example, a friendship tie between Facebook users either exists or it does not. In contrast, a *weighted* edge includes values associated with each edge that indicate the strength or frequency of a tie. For example, a weighted edge between two Facebook users may indicate the number of photo comments exchanged or the duration of a friendship. Weighted edges are often represented visually as thicker or darker lines or as more or less opaque lines. Including weighted data is preferable because they provide additional information about each tie. However, many social network analysis metrics (discussed later) are designed for unweighted networks. Fortunately, any weighted network can be converted to an unweighted one by choosing a cutoff point. For example, an unweighted edge could be shown between individuals who exchanged at least 10 email messages, with no edge between people who exchanged fewer than 10 messages.

3.2.4 Network Data Representations

Because network data differ from attribute data, there are different ways of representing it. With attribute data, it is common to create a data matrix where each row represents an individual and each column represents individuals' characteristics, behaviors, or answers to survey questions. A related approach can be used to represent relational data. Like attribute matrices, each row represents an individual in the network. However, unlike attribute matrices, each column also represents an individual as shown in Table 3.1.

Different types of edges can be represented in network matrices. Table 3.1 describes a directed network because not all connections are reciprocated. For example, Ann "points to" Bob as shown in row 1, but Bob

ADVANCED TOPIC
The Foundations of Graph Theory

Network analysis is rooted in the work of the mathematician Leonhard Euler who in 1736 studied whether a single path could be walked over the Seven Bridges of Königsberg that connected islands in the river Pregel (which flows through what was then Prussia and is now Kaliningrad in Russia) without crossing any bridge more than once. By reimagining the problem in terms of vertices and edges, he showed it is impossible to cross each bridge just once. Although the problem seems abstract, its solution led to the development of the mathematics of graph theory and, notably, hundreds of years later, the mathematical work of Paul Erdös and Alfréd Rényi on random graphs in the 1950s, an important theoretical development that allows for the generation of a graph from random processes. Social network analysis builds on these concepts and extends them to capture the nonrandom connections that occur among groups of people.

TABLE 3.1* A Network Represented as a Matrix

	Ann	Bob	Carol
Ann	0	1	1
Bob	0	0	0
Carol	1	0	0

**This network is a directed network, as it is not symmetrical (i.e., Ann points to Bob in row 1, but Bob doesn't point to Ann in row 2). It is a simple binary network: either a tie exists (value = 1) or not (value = 0).*

TABLE 3.2* A Network Represented as an Edge List

Vertex1	Vertex2
Ann	Bob
Ann	Carol
Carol	Ann

**Individuals in the Vertex1 column "point to" those in the Vertex2 column in this directed network. The network is implied to be a binary network. Additional columns could be used to describe each edge. For example, an Edge Weight column could be added with values representing the strength of various ties.*

does not "point to" Ann as shown in row 2. If it were an undirected network it would be a symmetric matrix; if Ann points to Bob then Bob must necessarily point to Ann. This network is a binary network because it only includes 1s and 0s, where a 1 indicates that there is a connection and a 0 indicates that there is no connection. Allowing additional values would create a weighted network. For example, the 1s could be replaced with the number of email messages sent to the other person. Finally, notice that the diagonal of the matrix connects each person with himself or herself. In this network, like most networks, the diagonal values are 0 indicating that a person does not "point to" herself. However, in some networks a "self-loop" connecting a person to herself can exist. For example, a person may send herself an email message as a reminder. Network matrices are powerful forms of representation that lend themselves to efficient mathematical manipulation for those inclined. However, they can also become quite large and challenging to navigate, particularly when networks have relatively few connections.

An alternate network representation is called an "edge list." Like its name suggests, it is simply a list of all edges in the network as shown in Table 3.2. This is the same network as shown in Table 3.1. Individuals in the Vertex1 column "point to" those in the Vertex2 column. Unless data describing the value of each edge are provided in additional columns, the network is implied to be a binary one. Self-loops are possible to represent in edge lists by having a row with the person's name repeated in both columns. Throughout this book, we will use edge lists instead of matrices.

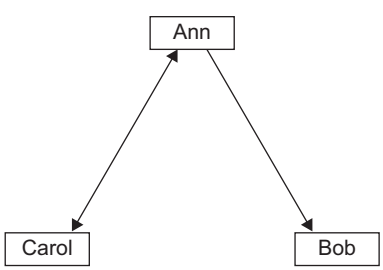

FIGURE 3.2 The directed, binary network described in Tables 3.1 and 3.2 represented as a network graph. Arrows indicate the direction of the connection (e.g., from Ann to Bob).

The final method for representing networks is through network graphs. Figure 3.2 is a network graph based on the data in Table 3.2. It makes immediately clear that the relationship between Ann and Carol is reciprocated (i.e., there are arrows on both sides of the line connecting them) and that there is no connection between Bob and Carol. Our earlier analysis of Figure 3.1, another network graph, demonstrates how network graphs can lead to insights that are hard to identify in tabular data, particularly when large networks are presented. However, many network graphs require significant preparation to assure that they are readable as described in Section 3.9.

3.3 TYPES OF NETWORKS

Social networks range in size from a handful of people to national and planetary populations. They also differ in the types of vertices they include, the nature of the edges that connect them, and the ways in which they are formed. In this section we introduce some of the distinctions that network scientists have identified to describe different types of networks. These distinctions affect the metrics and maps generated from them, as well as their interpretation.

3.3.1 Full, Partial, and Egocentric Networks

It is often useful to consider social networks from an individual member's point of view. Network analysts call the individual that is the focus of attention "ego" and the people he or she is connected to "alters." Some networks, called egocentric networks, only include individuals who are connected to a specified ego. For example, a network of your personal Facebook friends would be an egocentric network because you are, by definition, connected to all other vertices. Other egocentric networks and their associated "subgraphs" (see Chapter 6) may extend out from an ego, reaching not only friends, but also friends of friends. More generally, egocentric networks can extend out any number of "degrees" from ego. The basic "1-degree" ego network consists of the ego and their alters. The "1.5-degree" ego network extends the 1-degree network by including connections between all of the alters. For example, a Facebook 1.5 degree ego network would characterize which of your friends know each other (see Chapter 11). The "2-degree" ego network extends the 1.5-degree network by including all of the alters' own alters (i.e., friends of friends), some of whom may not be connected to ego. These three ego networks allow you to look at increasingly larger, but still "local" neighborhoods around a particular individual in a social network. Higher-degree networks (e.g., 2.5, 3) are feasible to create but not used as often in practice because they can quickly become intractable.

A full or complete network contains all the people or entities of interest and the connections among them. All egos are treated equally. A full network is often created and available when a single system, such as a social media platform, acts as a hub among a group of connected people or groups. For example, the Twitter network includes all users of the service and the connections between them. In practice, it is not always feasible (or particularly insightful) to analyze a full network. Instead, analysts create a partial network by selecting a sample or slice of the full network. For example, Figure 3.1 showed the slice of the Twitter network that included people who used the hashtag "#WIN09." This partial network was not egocentric. Rather it was topic centric. Other partial networks may be created to include a subgroup of users (e.g., all conference attendees), only people and connections that occurred within a specified time frame, or people who have certain characteristics (e.g., CEOs of Fortune 500 companies).

3.3.2 Unimodal, Multimodal, and Affiliation Networks

Up until this point we have only considered networks that connect the same type of entity. These standard networks are called *unimodal networks* because they include one type (i.e., mode) of vertex. They connect users to users or they connect documents to documents, but they don't include both users and documents. However, networks can include different types of vertices creating *multimodal networks*. For example, a network may connect users to discussion forums and blog posts they have commended on. Each vertex on the graph would represent a user, a forum, or a blog post, which could be visually distinguished by different colors or shapes. The SeriousEats Network discussed in Chapter 6 is an example. The rich sets of intersecting networks that form in social media environments include connections between people, photos, videos, messages, documents, groups, organizations, locations, and services. In many cases, these multimodal networks have to be transformed into simpler unimode networks to perform meaningful network analysis, as most network metrics (see Section 3.5) are designed for unimodal networks.

A common type of multimodal network is a *bimodal network* with exactly two types of vertices. Data for these networks often include individuals and some event, activity, or content with which they are affiliated, creating an *affiliation network*. For example, an affiliation network may connect users with wiki pages they edit. People are affiliated with pages. In this network, no two users would directly connect to each other. Likewise, no two wiki pages would directly connect to each other in this type of network.

Bimodal affiliation networks can be transformed into two separate unimodal networks: a user-to-user network and an affiliation-to-affiliation network (e.g., article–to–article network in a wiki) (see Chapter 6, Advanced Topic, Transforming Multimodal Affiliation Networks into Unimodal Networks for details). The user-to-user network connects people based on their links to one another. For example, in a wiki co-edit affiliation network Derek and Marc would be strongly connected because they both edit many of the same wiki

pages. The affiliation-to-affiliation network connects the affiliations based on the number of shared users. For example, a pair of wiki pages would be closely connected if many people edited both of the pages (see Chapter 15). More generally, this approach can be used to relate objects of all types (e.g., books, photos, and audio recordings) based on users' behaviors (e.g., purchasing or reading habits) and preferences (e.g., ratings). Affiliation networks are the raw material of many recommender systems that recommend items of interest, such as Amazon's "Customers Who Bought This Item Also Bought" feature. A network data structure can return results to queries like "people who linked to this document also linked to these documents" or "if you link to this document, you may want to link to these people."

3.3.3 Multiplex Networks

Although it is common for two people to be connected in many different ways (for example, by exchanging phone calls, emails, sharing group membership, and being married), most networks only include one type of connection or edge. However, it is possible to consider networks with multiple types of connections, called *multiplex networks*. For example, the Twitter network shown in Figure 3.1 includes three types of directed edges: following relationships, "reply to" relationships, and "mention" relationships. The graph could have uniquely represented each type of edge by using color, different edge types (e.g., dotted lines, solid lines), or edge labels (see Chapter 4). In the case of Figure 3.1, the type of edge was not deemed important, so the multiplex network data were condensed into a standard network that showed a single directed edge if one or more of the three types of connections were present. This strategy of combining multiple types of edges is a common one that allows for the use of network metrics, which are mostly based on standard networks.

3.4 THE NETWORK ANALYSIS RESEARCH AND PRACTITIONER LANDSCAPE

You can find network scientists in nearly every academic discipline and an increasing number of practitioner communities. Network concepts and techniques are now widely found throughout a range of disciplines including sociology, anthropology, communications, computer science, education, economics, physics, management, information science, medicine, political science, public health, psychology, biology, and the humanities. In the past several decades, social scientists have shown that network structures have an influence on health, work, and community. Getting a job, being promoted, catching an illness, adopting an innovation, and many more activities and processes have been explained in terms of social networks. Network structures are important in the biological sciences where research is focused on connections between metabolic and genetic processes. The shape and function of networks can have great consequences as ideas, genes, innovations, or pathogens diffuse through populations. Researchers now apply network theory and methods to understanding how Supreme Court decisions relate to previous cases, how the United States Senate votes (see Chapter 7), how epidemics spread within cities, and how characters in a novel relate to one another (see Chapter 7). Networks are formed from many physical processes and are echoed in a number of structures created inside information systems such as the collection of linked documents within the World Wide Web or an enterprise's collections of files. Information scientists use these links to identify high-quality web pages (e.g., Google's PageRank algorithm), or use the citations from research articles to identify high-impact articles and authors.

Network methods are diffusing beyond academic research, becoming an important tool for managing organizations, markets, and movements. Entrepreneurs apply network analysis techniques to understand how to leverage the powerful effects of word-of-mouth marketing as their customers spread news about their new products to one another. Many politicians recognize the potential power of a connected network of supporters who can be turned into contributors, volunteers, and voters. Engineers use network analysis to build more effective power grids, computer networks, and transportation systems. Law enforcement officers and lawyers analyze email networks to identify and defend potential criminals. And the intelligence community hunts down terrorists by looking at networks created by money trails and kinship. Having at least a basic understanding of network thinking and concepts is a core literacy of our time. Like statistics, network analysis has countless applications to a number of fields.

This book primarily focuses on social network analysis, a subfield of network sciences that focuses on networks that connect people or social units (i.e., organizations, teams) to one another (see Advanced Topic: Early Social Network Analysis). We are also interested in networks that connect human-generated content or artifacts together, such as web sites or cell phones.

ADVANCED TOPIC
Early Social Network Analysis

The social science roots of social network analysis can be found in the early 1800s in the work of the person credited with being the first sociologist, Auguste Comte, and later in the early 1900s in the work of the sociologist Georg Simmel. Both saw patterns of social ties as the main focus of sociology in contrast to the study of individuals and their attributes. Early in the nineteenth century, Comte defined society as more than simply a group of people. He argued that a population became a society only when people had influence on one another and considered the choices and interests of others as part of their own choices. Simmel echoed these ideas at the turn of the twentieth century, focusing social science on the study of how people come together and form groups and associations. These sociologists imagined society as composed of a web of relationships—more than a mass of individuals; they saw societies as networks of reciprocal influence.

The idea of connected actions linking people to one another has remained at the core of the social sciences, but efforts to create a systematic language to record social relationships started only in the twentieth century. Anthropologists studying the range of kinship systems they documented in fieldwork from around the world created symbol systems that are related to social network analysis. Their maps of who is related to whom were early forms of social networks focused on just the subset of social ties that are considered to be "family." The core concepts and methods of modern social network analysis date from the 1930s and the pioneering work of Jacob Moreno and his many collaborators. Researchers at New York University, Columbia, and Harvard created the first scholarly works featuring the distinctive core components of modern social network theory: measures, maps, and models. Moreno and his research partners created the first pictures of patterns of people and their partners, using visual maps with symbols that represented individuals with different types of lines connecting them to others that represented different kinds of relationships.

Moreno documented relationships among schoolchildren and the way an innovative behavior, running away, moved through chains of student connections. In 1934, Moreno [8] published "Who shall survive," which catalyzed work among a group of scholars who refined his approach and added critical mathematical elements that today are a standard part of network analysis. These approaches were applied to various settings, and revealed the key roles a few people played in their networks and often the presence of subgroups of distinct people. For example, in the 1930s, Davis et al. collected detailed records of observed attendance at 14 social events by 18 southern women, and the graph of that data revealed two distinct groups with minimal overlap [9]. Moreno, who developed sociometry and is often considered the founder of the sociogram, studied relationships among members of a football team and found patterns of friendship and animosity (see Fig 3.3) (as produced in Freeman [10]).

At Harvard in the 1930s, a group formed around W. Lloyd Warner and Elton Mayo to explore interpersonal relationship in workplaces. Early social network analysis work focused on connections in small work groups in industrial factory settings. For example, Roethlisberger and Dickson [11] studied the Western Electric Wiring room, documenting the ways individuals within a group worked with one another. As seen in Figure 3.4, some workers in the study emerged as the most connected, whereas others appeared as peripheral or isolated. Another data set was created that represented the relationships among 14 employees of the Western Electric Hawthorne Plant. Employees and two inspectors were observed, and each contact among them was coded. When employees played games with one another, argued, were openly friendly, confrontational, or helpful a note and tie was recorded. The result were six networks, which led to a seminal work by the Harvard sociologist George Homans [12] and later more mathematical work that focused on automatically finding clusters or groups within these data

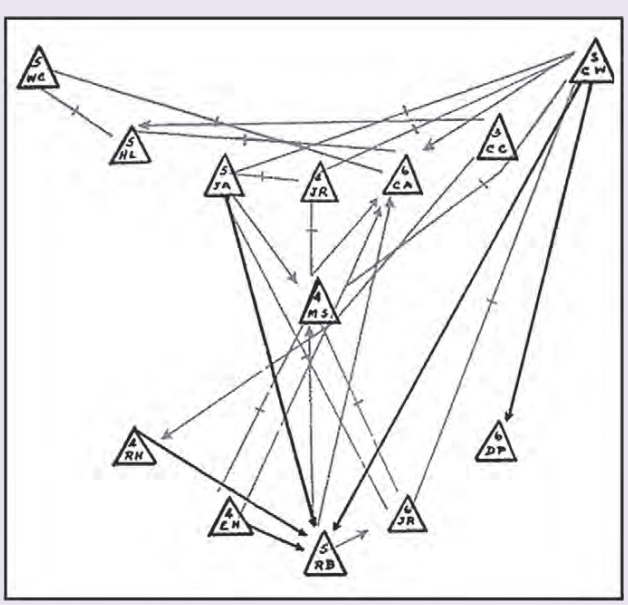

FIGURE 3.3 Jacob Moreno's early social network diagram of positive and negative relationships among members of a football team. Originally published in Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease Publishing Company.

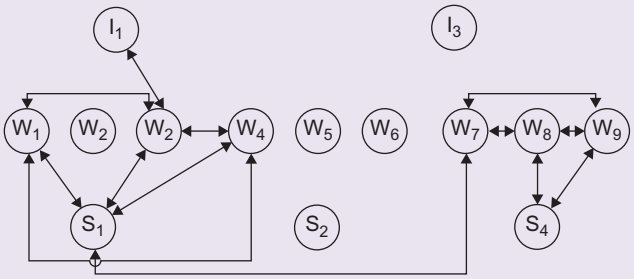

FIGURE 3.4 An early social network diagram of relationships among workers in a factory illustrates the positions different workers occupy within the workgroup. Originally published in Roethlisberger, F., and Dickson, W. (1939). *Management and the worker.* Cambridge, UK: Cambridge University Press.

sets [13]. In the 1950s, Nadel wrote about social roles and the social structures that define them [14]. He saw that the patterns of connections people had might be similar, even if they were connected to different people. These patterns, Nadel suggested, could be studied systematically, but in the 1950s the data and computational resources made that ambition a challenge.

Over time, Moreno's colleagues, including Paul Lazarsfeld, added key ingredients of the modern form of social network analysis: metrics and algorithms for calculating important network properties of the graph as a whole and for each individual in the graph (see Freeman [10] for details).

3.5 NETWORK ANALYSIS METRICS

Social scientists, physicists, computer scientists, and mathematicians have collaborated to create theories and algorithms for calculating novel measurements of social networks and the people and things that populate them. These quantitative *network metrics* allow analysts to systematically dissect the social world, creating a basis on which to compare networks, track changes in a network over time, and determine the relative position of individuals and clusters within a network.

Social network measures initially focused on simple counts of connections and became more sophisticated as concepts of density, centrality, structural holes, balance, and transitivity developed. Some metrics describe a network as a whole. For example, network density captures how highly connected vertices are by calculating the percentage of all possible connections that are realized. Other metrics are calculated for each vertex in a network. For example, centrality measures, of which there are many, capture how "important" (central) a vertex is within the network based on some objective criteria. Some people sit at the edge or periphery of their networks, whereas others are firmly at the center, connected to all the other most connected people. Even among a highly connected network, some pairs are not directly connected. When a third person bridges their connection, we can think of that person as a broker or connector. When that person is missing, we can think of a structural hole, a gap in which there is a missing connector. The following sections describe some of these metrics in more detail. Chapter 5 introduces some of the core metrics found in NodeXL through hands-on exercises.

ADVANCED TOPIC
Historical Obstacles to the Development of Network Analysis

Following the rapid development of the major elements of social network analysis in the 1930s there was a period of stagnation and neglect. For a variety of reasons, from Moreno's own personal and professional conflicts to the cost and lack of available network data sets and computing resources, social network analysis languished for decades.

The early social network literature was built on manually collected and processed data about social ties. Researchers would typically observe or survey population members, asking each to list those they came in contact with regularly for a variety of tasks and purposes. The prohibitive cost of this approach was a major limiting factor in the widespread application of social network analysis in enterprises and organizations. The recent explosion of computer-mediated social relationships and the associated drop in the costs of creating network data sets have made network approaches increasingly practical. As more details about our interactions and associations are tracked and captured by mobile devices and social media services, network analysis becomes increasingly useful.

Network analysis is computationally intensive: to generate many network metrics can require millions of calculations even when managing modest sized data sets. The recent explosion of computing power and the associated drop in costs have made network approaches increasingly practical, even if network methods remain among the most computationally intensive in use.

3.5.1 Aggregate Networks Metrics

A number of metrics describe entire networks. In some cases, a single network is broken into several disconnected pieces, called *components*. Some aggregate network metrics only work on networks where all of the vertices are connected in a single component, whereas others can be applied to entire networks even if they are split up. Here we describe just a few aggregate network metrics to give a flavor for what is possible, leaving a fuller discussion for Chapter 5.

Density is an aggregate network metric used to describe the level of interconnectedness of the vertices. Density is a count of the number of relationships observed to be present in a network divided by the total number of possible relationships that could be present. It is a quantitative way to capture important sociological ideas like cohesion, solidarity, and membership.

Centralization is an aggregate metric that characterizes the amount to which the network is centered on one or a few important nodes. Centralized networks have many edges that emanate from a few important vertices, whereas decentralized networks have little variation between the numbers of edges each vertex possesses.

Other metrics integrate attribute data with network data. For example, metrics that measure homophily look at the similarity of people who are connected. Studies typically show that people are connected to others who are similar to themselves on core attributes like income, education level, religious affiliation, and age.

3.5.2 Vertex-Specific Networks Metrics

Another set of metrics identifies individuals' positions within a network. Paramount among these is the set of centrality measures, which describe how a particular vertex can be said to be in the "middle" of a network. In the 1970s and 1980s, the sociologist Philip Bonacich developed a refined measure of centrality that took into consideration the different value a well-connected person can have in contrast to people with few connections. Network theorists noted that simply having many connections, called "degree centrality," was only one way to be "at the center" of things. A person with fewer connections might have more "important" connections than someone with more connections. One connection can be more important than another in different ways. Some are better because they bridge across otherwise separated portions of the network, whereas others are important because they connect to well-connected people. The following centrality metrics provide quantifiable measures for these concepts (see Chapter 5 for more details).

Degree Centrality

Degree centrality is a simple count of the total number of connections linked to a vertex. It can be thought of as a kind of popularity measure, but a crude one that does not recognize a difference between quantity and quality. Degree centrality does not differentiate between a link to the president of the United States and a link to a high school dropout. Degree is the measure of the total number of edges connected to a particular vertex. For directed networks, there are two measures of degree. In-degree is the number of connections that point inward at a vertex. Out-degree is the number of connections that originate at a vertex and point outward to other vertices.

Betweenness Centralities: Bridge Scores for Boundary Spanners

The notion of paths is central to the study of networks. Perhaps one of the most natural questions to ask about any two people in a network it is "How far apart are they?" This distance is measured simply: the distance between people who are not neighbors is measured by the smallest number of neighbor-to-neighbor hops from one to the other. For instance, people who are not your neighbors, but are your neighbors' neighbors, are a distance 2 from you, and so on. The shortest path between two people is called the "geodesic distance" and is used in many centrality metrics. For example, betweenness centrality is a measure of how often a given vertex lies on the shortest path between two other vertices. This can be thought of as a kind of "bridge" score, a measure of how much removing a person would disrupt the connections between other people in the network. The idea of brokering is often captured in the measure of betweenness centrality.

A structural hole is a missing bridge. Wherever two or more groups fail to connect, one can argue that there is a structural hole, a missing gap waiting to be filled. Burt provides compelling evidence that individuals who bridge structural holes are promoted faster than others [15]. Social network analysis has many strategic applications when people in an organization can analyze their position and the position of others. Managers and leaders can recognize gaps or disconnections within organizations and devote resources to traversing the divide. People may be able to apply social network analysis to identify locations in which a gap exists and elect to fill them, recognizing the value they can generate as broker between two otherwise separate groups.

Closeness Centrality: Distance Scores for Broadly Connected People

Closeness centrality takes a different perspective from the other network metrics, capturing the average

distance between a vertex and every other vertex in the network. Assuming that vertices can only pass messages to or influence their existing connections, a low closeness centrality means that a person is directly connected or "just a hop away" from most others in the network. In contrast, vertices in very peripheral locations may have high closeness centrality scores, indicating the number of hops or connections they need to take to connect to distant others in the network. Think of closeness, paradoxically, as a "distance" score. Some people are just a few miles from the big city, others must drive for hours: similarly, people with high "closeness" centrality scores have many miles or rather personal connections that they must travel to reach many other people in the network. Note that in some cases the inverse of the average distance to others in the network is used as a measure of closeness centrality. In that case, higher values indicate a more central position.

Eigenvector Centrality: Influence Scores for Strategically Connected People

Eigenvector centrality is a more sophisticated view of centrality: a person with few connections could have a very high eigenvector centrality if those few connections were themselves very well connected. Eigenvector centrality allows for connections to have a variable value, so that connecting to some vertices has more benefit than connecting to others. The PageRank algorithm used by Google's search engine is a variant of Eigenvector Centrality.

CLUSTERING COEFFICIENT: HOW CONNECTED ARE MY FRIENDS?

The clustering coefficient differs from measures of centrality. It is more akin to the aggregate density metric, but focused on egocentric networks. Specifically, the clustering coefficient is a measure of the density of a 1.5-degree egocentric network. When these connections are dense, the clustering coefficient is high. If your "friends" (alters) all know each other, you have a high clustering coefficient. If your "friends" (alters) don't know each other, then you have a low clustering coefficient. People have different measures for their clustering coefficient depending on the ways they cultivate connections to others and the environments they are in.

3.5.3 Clustering and Community Detection Algorithms

A network approach contrasts with those that presume the existence and boundaries of groups. In a network perspective, people occupy many relationships and are potentially members in many groups and less defined clusters. Defining exact boundaries in networks may be difficult, reflecting the reality of multiple and shifting memberships. From a network perspective, a group is a collection of vertices that are more connected to one another than they are to others. Relatively more cohesive or densely connected sets of vertices form regions, also called clusters, that may reflect the existence of groups without regard to whether they are officially recognized or even if members recognize their connections to one another. A rapidly growing body of research describes clustering algorithms, also called community detection algorithms, that automatically identify these clusters based on networks structures. We discuss these in more detail in Chapter 7.

3.5.4 Structures, Network Motifs, and Social Roles

Two people within a network may sometimes share a pattern of connection to other people, even if they do not connect to the same people. Certain professions have distinct patterns of connections, either linking with many others (real estate agents, and other retail professionals) or few (reclusive authors and artists, peripheral workers, and other people focused on things rather than people). In addition to the number of connections, some people share the pattern of connections among the people they connect. In some cases people are connected to people who are strangers to one another, in other cases a group may be densely connected to one another. These secondary patterns of connection are a distinctive feature of network analysis approaches: networks are as much about the attributes and patterns of connection among neighbors as they are about the attributes and connections of any individual.

Social roles are complex cultural and structural features of social life. An example social role like "father" is explicitly recognized in society, has a wide set of culturally shared meanings and expectations, is associated with particular goals and interests, and is partly defined by the content and structure of actions directed toward other distinctive role holders. Although social roles may not be as clearly defined or explicitly recognized by all the actors in a given social setting, they have identifiable content, behavioral, and structural features.

Studies of social media have illustrated the ways contributors create distinctive network patterns that reflect their role or status within the community (e.g., Welser, Gleave, and Smith [16]). These patterns are evidence of specialization of behavior in these social spaces. An example of a role in a social media space is the "answer person" who disproportionately provides the answers to questions asked in message board environments, "discussion people" who engage in extended exchanges of

messages in large and populous threaded discussions, "discussion starters" who demonstrate influence over the topics discussed by the "discussion people," "influential" people who are well connected to others who are more highly connected than they are, and boundary spanners who bridge between unconnected subgroups. These roles are described in greater detail in Chapter 9, devoted to email lists and discussion groups.

ADVANCED TOPIC

A Renaissance of Network Research and Data

Since the 1960s, network analysis has blossomed. New research and methods have flourished and social networking has developed a new prominence in mainstream culture. Despite early challenges, in the past several decades a healthy and growing subfield has reemerged around social network analysis. New network tools and concepts have been created and applied to a wide and growing range of domains. Mathematical sociology has developed as a major subdiscipline in the social sciences, dedicated to finding elegant descriptions of complex social phenomena. Starting 80 years ago with simple hand-drawn charts and diagrams that described small groups of people and their connections, network science concepts, methods, and tools are used today to calculate a range of measures that describe the shape, structure, and dynamics of potentially multimillion or billion vertex networks. New methods have been developed for automatically organizing and displaying visualizations of the links among large populations. This combination of structural models, visualizations, and metrics forms the key features of modern social network analysis.

In the late 1960s, Stanley Milgram explored the idea of small world networks in a study that came to be referred to as "Six Degrees of Separation" [17], which later inspired the 1990 John Guare play and 1993 movie of the same name. The study explored the question of how connected any two people selected at random might be. Milgram sent a collection of letters to people around the country asking them to send the message to someone they knew who could move their letter closer to the target, a stock broker in Massachusetts. On average, the letters took six steps to arrive at their destination. The "six degrees" or steps suggested that even in large networks where most people are not directly connected, people can be reached from every other person through a small number of steps.

Sampson's study in the late 1960s of relationships among members of a residential monastery captured social network data during an event in which several members were expelled or chose to leave [18]. A series of social network data sets were collected by asking participants about who they liked and spent time with. Social network analysis of this data allowed Sampson to identify the future lines of division among the members of the network. The idea that members of a network can be grouped based on how densely they are connected is an important concept in network analysis. These groups can be important divisions with consequences for the future of the network. For example, a notable study by Zachary in the 1970s mapped the structure of a Karate club based on affinities and connections between students and teachers. These maps predicted the ways the club eventually split when a new teacher, in conflict with the owner, left the studio and took many students with him [19].

The sociologist Barry Wellman demonstrated in the 1970s that real-world communities are composed of interlocking social networks of specialized relationships that changed dramatically in composition over a period of years.[2] He proposed that society was now characterized by networked individualism in contrast to the group memberships and identities of prior periods. Rather than defining oneself in professional or political terms, people create personal networks in which they occupy distinct locations and roles. He later applied these techniques to study online networks [20]. In 1977, Wellman founded a social network analysis professional association, the International Network for Social Network Analysis (INSNA). INSNA now has more than a thousand members, many of whom have gathered for more than 20 years for an annual conference ("Sunbelt") on social network analysis research.[3] Journals and publications devoted to social network analysis include *Social Networks*, *Connections*, and the *Journal of Social Structure*. Social network data, methods, and visualizations appear across a much wider spectrum of journals and conference publications.

[2]www.chass.utoronto.ca/~wellman/vita/index.html

[3]www.insna.org

In the early 1970s, the sociologist Mark Granovetter did research on the employment market, looking at how people discovered new job opportunities. He observed that, in contrast to the view held by classical economics, people were not freely floating independent actors in the labor market. They were embedded in a set of different relationships with particular people. Granovetter found that job news passed through connections that were not the closest and most intense relationships [21]. A person's "weak ties" brought news from distant parts of the social network to which "strong ties" did not have access because they occupied such a similar place in the network as the job seeker. Thus weak ties proved particularly useful for finding novel information, such as information about job prospects. Because weak ties were less intense, they were also less costly to maintain in terms of time and attention. As a result, it is possible to have many weak ties but only a few strong ties.

Armed with new network metrics and the means to calculate them, network analysts have focused on a variety of data sources and questions. Social networks have been applied to historical studies using records of investments, marriages, and memberships in elected positions. In the 1400s in the city of Florence, the Medici and Strozzi families struggled for domination. These families, like many others, were locked in political struggles. In the 1970s, John Padgett collected records of the social relations among Renaissance Florentine families that he extracted from historical documents. Families were often connected through a variety of ties, relations, and business connections. A data set was created that represented the financial loans, credits and joint partnerships, and marriages that bound families to one another. The resulting data set included information about each family as well as their links to others. Each family had a value representing its net wealth in the year 1427, the number of seats it held in the local government between the years 1282 and 1344, and the number of business or marriage ties among the population of 116 families. Analyzing these data, Padget found that the Medici held great power because, he argued, they sat at the center of business and family networks, brokering connections that no other family could equal [22, 23].

A more modern version of the study of historical Florentine politics can be found in the study of interlocking directorships in modern corporations. Many corporations and other institutions have a board of directors, some of whom serve on more than one board. When board members serve on two or more boards, they link those corporations and, in aggregate, create interlocking directorships that combine to form even larger meta-institutions. By building on research on interlocking directorships in U.S. corporations [24, 25], modern web sites like "They Rule" provide an interactive map that displays the common links between major corporations.[4]

In 1992, Robin Dunbar famously argued that people have an innate ability to handle a number of social relationships but not an endless number of them. Remembering people's names may have a biological limit as our brains evolved over long periods in which there were rarely more than a few hundred people within any region, group, or tribe. The number 150 has been loosely associated with the idea of a "Dunbar" number, an upper limit on the number of relationships a person can normally manage.[5] This number can be expanded with augmentation, through analog technologies like diaries, address books, and the "filo-fax." More recently, social media tools like Facebook and email contact lists extend our ability to maintain more relationships. These additional relationships can be said to be "weaker" than the core 150 "organic" relationships, but as Granovetter has shown, weak ties can collectively be of enormous value.

Business Applications of Social Network Analysis

Social network analysis has historically been an academic endeavor, but as network analysis tools and data sets become more available, pioneering businesses are applying it to help manage business challenges, gain insight into markets and communities, and build more robust industry relationships. For example, the work of Rob Cross and the Network Roundtable focuses on several practical applications of social network analysis for corporations and other large organizations, highlighting differences between healthy and underperforming divisions and the value of organization spanning connections [26, 27]. Others apply network analysis to the improvement of corporate structures and processes [28]. In the early 1990s, Monge and Contractor [29] documented the many forms of social network patterns that emerge inside of organizations and institutions.

Social networks have been shown to have a significant influence on the adoption of new technologies or social practices. The sociologist Everett Rogers described the concept of the "diffusion of innovations," arguing that

[4]www.theyrule.net

[5]www.lifewithalacrity.com/2004/03/the_dunbar_numb.html

(*Continued*)

ADVANCED TOPIC (*Continued*)

people with particular patterns of connections to others played pivotal roles in the success or failure of a new idea or message being adopted or distributed through the network [30]. Networks with different patterns of connection have different properties in terms of how they propagate a new message, rumor, or product and how they resist being dissolved when vertices are removed from the graph. These observations have significant implications for interventions into disease and rumor propagation and the cultivation of innovation [31].

Networks play an important role in e-commerce where collaborative filtering powers the familiar list of "books that people who liked this book also liked." Businesses are also interested in learning the requirements of viral marketing. We will discuss diffusion and marketing in more detail in the discussion of Twitter in Chapter 10, but for now the important thing to know is that diffusion can often lead to "cascades" where an unknown, even marginal idea can spread rapidly throughout the entire network and become the norm. Memes are a commonly-cited example of contagion, as are viral products, such as viral videos on YouTube that go from dozens to millions of viewers in a few months or even weeks.

3.6 SOCIAL NETWORKS IN THE ERA OF ABUNDANT COMPUTATION

The widespread adoption of networked communication technologies has significantly expanded the population of people who are both aware of network concepts and interested in network data. Although the idea of networks of connections of people spanning societies and nations was once esoteric, today many people actively manage an explicit social network of Internet friends, contacts, buddies, associates, and addresses that compose their family, social, professional, and civic lives. Email messages forwarded from person to person have become a common and visible example of the ways information passes through networks of connected people. The notion of "friends of friends" is now easy to illustrate in the features of Internet social media applications like Facebook, MySpace, and LinkedIn that provide explicitly named "social networking" services. Viral videos and chain emails illustrate the way word of mouth has moved into computer-mediated communication channels. The idea of "six degrees of separation" has moved from the offices of Harvard sociologists to become the dramatic premise of a Broadway play to now appear as an expected feature of services that allow people to browse and connect to their friend's friends.

As network concepts have entered everyday life, the previously less visible ties and connections that have always woven people together into relationships, cliques, clusters, groups, teams, partnerships, clans, tribes, coalitions, companies, institutions, organizations, nations, and populations have become more apparent. Patterns of sharing information, investments, personal time, and attention have always generated network structures, but only recently have these linkages been made plainly visible to a broad population. In the past few decades, the network approach to thinking about the world has expanded beyond the core population of researchers, analysts, and practitioners who have applied social network methods and perspectives to understand their businesses, communities, markets, and disciplines. Today, because many of us manage many aspects of our social relationships through a visibly computer-networked social world, it is useful for many more people to develop a language and literacy in the ways networks can be described, analyzed, and visualized. Visualizing and analyzing a social network is an increasingly common personal or business interest. The science of networks is a growing topic of interest and attention, with a growing number of courses for graduates and undergraduates and even becoming the topic of a television documentary.[6]

The availability of cheaper computing resources and network data sets has enabled a new generation of researchers access to studies of the structures of social relationships at vastly larger scale and detail. Since the late 1960s, as computing resources and network data sets have grown in availability and dropped in cost, researchers began developing tools and concepts that enabled a wider and more sophisticated application of social network analysis.

[6]"Connected: The Power of Six Degrees," http://ivl.slis.indiana.edu/km/movies/2008-talas-connected.mov

ADVANCED TOPIC

Social Network Analysis Research Meets the Web

As access to electronic networks grew in the 1970s, academic and professional discussions and collaborations began to take place through them. Systems to support the exchange of messages and the growth of discussions and even decisions became a major focus of systems development and the focus of study itself. Freeman and Freeman [32] collected data from the records of the Electronic Information Exchange System (EIES) that itself hosted a discussion among social network researchers. Two relations were recorded: the number of messages sent and acquaintanceship. These systems became the focus of the first systematic research into naturally occurring social media. Even before the Internet, early computer network applications supported the creation of exchanges, discussions, and therefore social networks, built by reply connections among authors.

Early proprietary systems evolved into the public World Wide Web. In the 1990s, the computer scientist Jon Kleinberg identified the patterns of links between high-quality web pages, an algorithm that went on to inspire Stanford graduate students who founded the Google corporation. Kleinberg described different locations in a population of linked documents on the World Wide Web. On the World Wide Web, a document or page can link to another page, forming a complex network of related documents. Some documents contain many pointers to other documents, whereas others have many documents point at them. These hubs and authorities defined two broad classes of web pages that offered a path to identifying high-quality content. Links from one page to another are considered to be indicators of value. Refinements of the HITS algorithm made use of eigenvector centralities to implement the page rank algorithm that is the core of the Google web search ranking method [33].

Network researchers studying social networks and the Internet found that empirical networks often exhibit "small-world properties": most nodes are not neighbors with each other, but nodes can be reached from every other node in a small number of hops. In the late 1990s, the physicist/sociologist Duncan Watts, working with the mathematician Steven Strogatz, created mathematical models of "small world" networks and contrasted them with purely random networks such as those proposed by Erdos and Renyi [34]. Their model captured the natural properties of social networks far better than those that assumed a purely random or normal distribution of links. Although most connections are to others who are local, a few connections importantly can jump far from the individual. Many of our friends are likely to live or work near us, but a few may be very far away. These relatively rare far-reaching links can dramatically change the properties of the network, making the widespread transmission of messages much easier. This model significantly improved on earlier models of network growth and structure, better approximating the observed structure of naturally occurring social networks. Later researchers have built upon their work to devise models that generate "small world" networks that more closely match empirical networks, helping us to understand how networks may have become the way they are. For example, Barabasi and Albert have developed a family of models of preferential attachment that can generate "scale-free" networks, which are a common feature of social networks [35]. Scale free networks have a power law degree distribution, meaning that there are a few key hubs in a network and many poorly connected vertices. While none of these models perfectly predict social networks, they provide a method for systematically comparing networks and focus attention on the processes that may have led to the characteristics that we see in networks.

In the past few years, researchers have begun to study large web-based networks. For example, Leskovec and Horvitz calculated metrics for a graph that includes more than 300 million users of the Microsoft Messenger service [36]. Each user typically had one or more "buddies" with whom he or she might send one or more messages and receive some in return. Buddies often listed their locations, allowing these linkages to be aggregated into a complex map of the world and the flow of conversation around it. Others have reported on the hyperlink network created by web pages hyperlinking to other web pages (e.g., Park and Thelwall [37]). A number of studies have examined the blog network. For example, Adamic and Adar [38] showed how political blogs are divided into two clear clusters with minimal overlap that represent the left and right political populations in the United States. More recently, Kelly and Etling mapped Iran's blogosphere, identifying more than 20 subcommunities of bloggers who wrote in Farsi for an Iranian audience.[7]

Another line of research has focused on visualizing social networks. A representative paper by Heer and Boyd [39] described a tool called Vizster that allowed users to navigate through their friends from a social networking site to explore social connections.

[7]http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public

3.7 THE ERA OF ABUNDANT SOCIAL NETWORKS: FROM THE DESKTOP TO YOUR POCKET

We now live in a new era of network data abundance. Network data collection was once a time-consuming and laborious process that yielded small data sets at great cost. Observations, surveys and interviews took many days or weeks to perform, could not be repeated frequently, required many people to produce, and often yielded low rates of participation with inherent biases and errors. Asking people about their relationships with others continues to have benefits and offers unique sources of insight, but people have been shown to be a poor source of accurate information as bias and faulty memory warp what people report about who they know and with whom they interact. The challenge of creating a data set that spanned long periods or large numbers of people or contained records of many events proved insurmountable using traditional methods.

Today, interactions between people increasingly take place through computing systems. Users create many types of networks in a machine-readable form each day as our interactions are documented in a computer. When we use these communication tools, databases are created and maintained with records and log files that document the details of the time, place, and participants of each interaction, whether via computers or telephones and even televisions. These event logs describe many different kinds of connection but share a common structure in which one person or entity is linked to another by some relationship.

The creation of these machine-readable network data sets mean that long periods of time or large populations connected by many events can now be studied using widely available computing equipment and data sources.

Like a jump from Galileo's handmade telescope to the orbiting *Hubble*, network science has made a vast leap in scale and scope as we create a digitally networked world around ourselves.

As the historical drought of social network data has ended with a flood of sources of network data, the challenge has been to rapidly develop the tools and concepts needed to process and analyze them. Technical methods for building multiterabyte databases have shifted to the even vaster task of managing petabytes of data. New methods of harnessing thousands and even millions of computers in parallel have been driven by the growing need to manage vast data stores growing from the web. The challenge is likely to grow steeper as new sources of network data come pouring out off an emerging class of sensor-rich devices that record vast streams of data from millions to billions of people, devices, and locations. The early wave of this surge of data can be seen in new sources of data from everyday life that are being captured and recorded with mobile devices, creating a new stream of archival material that is richer than all but the most obsessively observed biographies. It has become common in recent years that the most timely and well-placed photographs and video recordings have come from everyday individuals with phones and computers rather than from news photographers and reporters.

The coming wave of mobile technologies is likely to deepen this trend, with new ways for phones or other devices to capture information about their users and the world around them. Research projects like SenseCam from Microsoft Research, which captures a continuous stream of photographs and temperature and motion data, are now becoming products,[8] and services like nTag, Spotme, Loopt, Foursquare, and Google Latitude using devices like iPhones and Droids are weaving location into every application (see Chapter 2).

As phones are increasingly able to notice each other, a new set of mobile social software applications are becoming possible, as evidenced by new services such as Bump.[9]

A service like SenseNetworks[10] is a good example of a mobile data collection, analysis, and presentation service. Other services and products like CureTogether[11] and FitBit[12] are examples of social movements that are enabled by web applications integrated with devices that provide self-monitoring medical tracking. These communities overlap with the trail-based exercise communities of runners, bikers, skaters, hikers, and skiers, some with artistic inclinations (they hike in paths that resemble drawings when seen on a map). A new wave of devices is emerging that extensively quantifies your "self" and "others," allowing you to perhaps swap sensor data with other people. The result could be an aggregated map of the heath and environmental conditions of the planet, not unlike early examples of collectively authored road maps of whole nations accomplished by the Open Street Map project.[13]

[8]www.nytimes.com/2010/03/09/health/09memory.html
[9]http://bu.mp
[10]www.sensenetworks.com
[11]www.curetogether.com
[12]www.fitbit.com
[13]www.openstreetmap.org/

3.8 TOOLS FOR NETWORK ANALYSIS

The growth of interest in network analysis has been dramatic, but until recently the development of social network analysis tools has lagged, and they remained a challenge to use for many people. Applying network approaches has been traditionally a challenge that involved much more than simply mastering a new set of concepts and ideas that focus on relationships and patterns. Network data have traditionally been difficult to create and collect, and the tools for analyzing and visualizing networks have demanded significant technical skill and often mastery of programming languages. Many tools that exist to support network analysis demand significant commitment to learn and master. Existing network tools that are relatively easier to use have typically lacked support for easily importing social media network data. In the past few years, many network analysis projects and research papers have focused on computer-mediated networks of people, documents, and systems. Only recently have new tools made it simpler for people to extract data from major social media network sources and to perform a basic network analysis workflow without requiring programming skills or using a command line interface.

Social media network data collection, scrubbing, analysis, and display tasks have historically required a remarkable collection of tools and skills. A great example of the variety of tools that can be used in concert to extract, analyze, and display social media networks can be found on Drew Conway's blog.[14] This is a powerful set of tools for those who can master the demands of python or other programming languages and the application programmer interfaces (API) that give sophisticated users special access to resources. In contrast, this book focuses on a single tool designed for nonprogrammers, NodeXL, because of its relative ease of use, support for rich visuals and analytics, and integration with the ubiquitous Excel spreadsheet software. The python path is certainly the high road for experts and those with demanding volumes of data or esoteric data analysis requirements. But for the noncoding user, NodeXL may be one of the easiest ways to both manipulate network graphs and get graphs from a variety of social media sources. A detailed step-by-step guide to the core features of NodeXL can be found in Part II of the book.

3.9 NODE-LINK DIAGRAMS: VISUALLY MAPPING SOCIAL NETWORKS

One of the key elements that characterizes modern social network analysis is the use of visualizations of complex networks. Compared to staring at edge lists or network matrices (see Section 3.2.4), looking at a network graph can provide an overview of the structure of the network, calling out cliques, clusters, communities, and key participants. It could be said that a graph is worth a thousand ties. Not only can network visualizations inspire understanding and insights, they can also be appealing and even beautiful. They can serve as persuasive tools that demonstrate important points about networks. The ability to map attribute data and network metric scores to visual properties of the vertices and edges (see Chapters 4 and 5) makes them particularly powerful.

However, network visualizations are often as frustrating as they are appealing. Network graphs can rapidly get too dense and large to make out any meaningful patterns as illustrated in Figure 3.5. Many obstacles like vertex occlusions and edge crossings make creating well-organized and readable network graphs challenging. There is an upper limit on the numbers of vertices and edges that can be displayed in a bounded set of pixels; typically only a few hundred or thousand vertices can be meaningfully and distinctly represented on average-sized computer screens. In his appeal for better-quality network visualization, Shneiderman [40] has suggested that we aspire to reach the worthy but not always attainable goal of "netviz nirvana" in which the following goals are proposed:

- Every vertex is visible.
- Every vertex's degree is countable (i.e., the number of connections that start or end at that vertex).
- Every edge can be followed from source to destination.
- Clusters and outliers are identifiable.

To approach netviz nirvana, careful preparation, layout, and filtering techniques must be used. In practice, network visualizations often fall far from the mark. However, the graphs shown throughout this book illustrate the value of carefully crafting network graphs. We hope they will inspire network analysts to take the care needed to create substantive, understandable, and aesthetically pleasing graphs.

3.10 COMMON NETWORK ANALYSIS QUESTIONS APPLIED TO SOCIAL MEDIA

Once a set of social media networks has been constructed and social network measurements have been calculated, the resulting data set can be used for many applications. For example, network data sets can be used to create reports about community health, comparisons of subgroups, and identification of important

[14]www.drewconway.com/zia/?p=204

individuals, as well as in applications that rank, sort, compare, and search for content and experts.

The value of a social network approach is the ability to ask and answer questions that are not available to other methods. This means focusing on relationships. Although analysts, marketers, and administrators often track social media participation statistics, they rarely consider relationships. Traditional participation statistics can provide important insights about the engagement of a community, but can say little about the connections between community members. Network analysis can help explain important social phenomena such as group formation, group cohesion, social roles, personal influence, and overall community health. Combining traditional participation metrics with network metrics provides the best of both worlds and allows you to answer important questions such as the following:

- What kinds of social roles are being performed within a social media repository? Does a community have enough people filling the important roles?
- Which individuals play important social roles within a group or collection? Who would make a good administrator based on that person's network position?
- What subgroups exist? Do connections between subgroups exist? Who plays the bridge roles that connect otherwise unconnected groups?
- How do new ideas propagate through a network? Who are the influencers?
- How do the overall structures of a social network change after a particular event (e.g., a company social, a round of new hires or layoffs)?

3.11 PRACTITIONER'S SUMMARY

The opportunities for practitioners to apply network analysis to contemporary business, community management, political influence, and team collaboration have dramatically increased in recent years. The once esoteric

FIGURE 3.5 A medium-sized node-link network diagram visualization of Twitter users linked by patterns of following. This sized graph illustrates many issues with a network graph containing more than a few dozen vertices. Many vertices sit on or overlap with other vertices. The number of edges associated with some vertices is impossible to count, whereas other edges cannot be traced from source to destination. Improvements to network layout are an active area of research.

concepts and metrics of network analysis have become talk show and airport lounge topics. The difficulties in collecting and analyzing network data have been dramatically reduced by powerful database methods and well-designed network analysis and visualization tools. There is still a lot of work to be done, but practitioners now have the potential to make more effective decisions based on network analyses of their own data conducted in a few hours, rather than a few months.

Learning the concepts and tools is a necessary first step, but the payoffs are large. The growing industry of social media and networking consultants complemented by a vast array of books plus informative web sites, online seminars, and Wikipedia pages, makes the necessary training widely available. At the same time, network analysis methods are spreading through university curricula rapidly and filtering into high school courses.

Attending public seminars and professional conferences provides other means to acquire skills and make valuable connections. Your first steps may be a struggle, but we hope that with each step the processes become smoother and the professional benefits larger.

3.12 RESEARCHER'S AGENDA

The research progress on network analysis has been dramatic in the past few decades, transforming an exotic research topic into a thriving research community in academia, government, and industry. The existing metrics, clustering, and layout algorithms are stabilizing, but innovative approaches are still emerging to trigger bursts of new research. As practitioner pressure builds to apply network analysis to ever larger data sets, researchers have developed remarkably more efficient algorithms, while hardware developers have produced powerful graphics processors (based on gaming computers), huge arrays of computers, and scalable cloud computing services. Meanwhile, new social media services generate more relational data than ever before, ushering in a golden era of social science research on human relationships and collaboration.

The algorithms and hardware provide the platforms, but the concomitant development of vastly improved user interfaces for network analysis has begun to enlarge the community of users from the dedicated sociologists who are also programmers to the broad segment of business analysts who use spreadsheets or simplified web-based tools. Packaging the complex processes of frequently applied network analyses into a few clicks is the next challenge in many fields, thereby inspiring other researchers and developers to simplify the processes even further, while increasing the power offered to users. The best is yet to come.

References and Resources

[1] J.P. Scott, Social Network Analysis: A Handbook, Sage, Thousand Oaks, CA, 2000.

[2] D. Easley, J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, Cambridge, UK, 2010.

[3] M. Newman, A.-L. Barabasi, D.J. Watts (Eds.), The Structure and Dynamics of Networks, Princeton University Press, Princeton, NJ, 2006.

[4] P. Carrington, J. Scott, S. Wasserman (Eds.), Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences), Cambridge University Press, Cambridge, UK, 2005.

[5] W. de Nooy, A. Mrvar, V. Batageli, Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences), Cambridge University Press, Cambridge, UK, 2005.

[6] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences), Cambridge University Press, Cambridge, UK, 1994.

[7] B. Wellman, Structural analysis, in: B. Wellman, S.D. Berkowitz (Eds.), Social Structures, Cambridge University Press, Cambridge, UK, 1988, pp. 19–61.

[8] J.L. Moreno, Who shall survive? A new approach to the problem of human interrelations, Nervous and Mental Disease Publishing Co., Washington, 1934.

[9] A. Davis, B.B. Gardner, M.R. Gardner, Deep South: A social Anthropological Study of Caste and Class, University of Chicago Press, Chicago, Ill., 1941.

[10] L.C. Freeman, The Development of Social Network Analysis: A Study in the Sociology of Science, BookSurge, LLC, North Charleston, SC, 2004.

[11] F. Roethlisberger, W. Dickson, Management and the Worker, Cambridge University Press, Cambridge, UK, 1939.

[12] G. Homans, The Human Group, Harcourt-Brace, New York, 1950.

[13] R. Breiger, S. Boorman, P. Arabie, An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling, J Math Psychol 12 (1975) 328–383.

[14] S.F. Nadel, The Theory of Social Structure, Cohen & West, London, 1957.

[15] R. Burt, Brokerage and Closure: An Introduction to Social Capital, Oxford University Press, Oxford, 2005.

[16] H. Welser, E. Gleave, M. Smith, Visualizing the signatures of social roles in online discussion groups, J. Soc. Struct. 8 (2) (2007).

[17] S. Milgram, The small world problem, Psychology Today 2 (1967) 60–67.

[18] S. Sampson, Crisis in a cloister. Unpublished doctoral dissertation, Cornell University, 1969.

[19] W. Zachary, An information flow model for conflict and fission in small groups, J Anthropol Res 33 (1977) 452–473.

[20] B. Wellman, An electronic group is virtually a social network, in: Kiesler Sara (Ed.), Culture of the Internet, Lawrence Erlbaum, Mahwah, NJ, 1997.

[21] M. Granovetter, The strength of weak ties, Am J Sociol 78 (6) (1973) 1360–1380.

[22] J. Padgett, C. Ansell, Robust action and the rise of the medici, 1400-1434, Am. J. Sociol. 98 (6) (1993) 1259–1319.

[23] D. Kent, The Rise of the Medici: Faction in Florence, 1426-1434, Oxford University Press, Oxford, 1978.

[24] M. Mizruchi, L.B. Stearns, A longitudinal study of the formation of interlocking directorates, Adm. Sci. Q 33 (2) (1988) 194–210.

[25] B. Mintz, M. Schwartz, The Power Structure of American Business, University of Chicago Press, Chicago, 1985.

[26] R. Cross, R.J. Thomas, Driving Results through Social Networks: How Top Organizations Leverage Networks for Performance and Growth, Jossey-Bass, San Francisco, CA, 2009.
[27] R. Cross, R.J. Thomas, Driving Results through Social Networks: How Top Organizations Leverage Networks for Performance and Growth, John Wiley & Sons, San Francisco, CA, 2009.
[28] M. Kilduff, W. Tsai, Social Networks and Organizations, Sage, Thousand Oaks, CA, 2003.
[29] P.R. Monge, N. Contractor, Theories of Communication Networks, Oxford University Press, New York, 2003.
[30] D.E.M. Rogers, Diffusion of Innovations, fifth ed., Simon and Schuster, New York, 2003.
[31] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.
[32] L.C. Freeman, Centrality in social networks: conceptual clarification, Social Networks 1 (1979) 35–41.
[33] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, in: Proc. 7th World-Wide Web Conference (WWW7), BrisBane, Australia, April 1998.
[34] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (4 June 1998) 440–442.
[35] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (15 October 1999) 509–512.
[36] J. Leskovec, E. Horvitz, Planetary-scale views on a large instant-messaging network, in: Proc. 17th international Conference on World Wide Web (Beijing, China, April 21–25, 2008). WWW '08. ACM, New York, NY, 2008, 915–924.
[37] H.W. Park, M. Thelwall, Hyperlink analyses of the world wide web: a review, Journal of Computer Mediated Communication 8 (4) (July 2003).
[38] L.A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25 (3) (July) (2003) 211–230.
[39] J. Heer, D. Boyd, Vizster: Visualizing Online Social Networks, in: Proc. 2005 IEEE Symposium on Information Visualization (October 23-25, 2005), INFOVIS. IEEE Computer Society, Washington, DC, 2005.
[40] B. Shneiderman, A. Aris, Network visualization with semantic substrates, IEEE Symposium on Information Visualization and IEEE Trans, Visualization and Computer Graphics 12 (5) (2006) 733–740.

Additional Resources

Barabasi, A.L. (2003). *Linked: How everything is connected to everything else and what it means*. New York: Penguin Group.
Bonacich, P. (1987). Power and centrality: A family of measures. *The American Journal of Sociology*, *92*(5), 1170–1182.
Borgatti, S., Mehra, A., Brass, D., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895.
Brandes, U., & Erlebach, T. (Eds.), (2005). *Network analysis: Methodological foundations*. Berlin, Heidelberg: Springer-Verlag.
Buchanan, M. (2002). *Nexus: Small worlds and the groundbreaking theory of networks*. New York, NY: Norton.
Burt, R. (1995). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.
Burt, R.S. (1992). *Structural holes*. Cambridge, MA: Harvard University Press.
Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library and Information Science Research*, *18*(4), 323–342.
Johnson, S. (2002). *Emergence: The connected lives of ants, brains, cities, and software*. London, UK: Penguin.
Knoke, D., & Yang, S. (2007). *Social network analysis (Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA: Sage.
Nooy, W., De, Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with pajek*. Cambridge, UK: Cambridge University Press.
Watts, D. (1999). *Small worlds*. Princeton, NJ: Princeton University Press.
Watts, D. (2003). *Six degrees*. New York: Norton.
Wellman, B., & Berkowitz, S.D. (1988). *Social structures: A network approach*. Cambridge, UK: Cambridge University Press.