The Little Book of Network Analysis

Jerry Scripps

October 22, 2016

Contents

Preface

This book is written specifically for CIS 310 a course at Grand Valley State University about the mathematical structure and machine considerations behind networks. It is intended to be background reading material where I was not able to find appropriate, available resources. The course is a general education course so the reader is not assumed to have any mathematical, computer, or statistical training beyond that of the typical undergraduate student.

Chapter 1

Introduction

One hears about networks frequently in the modern world. The word is used to describe a wide range of different complex systems in electronics, communications, biology, business and many other areas. These networks are often large, complicated systems that are difficult to understand. In this book, a network is an abstract model that we use to help us understand the relationships without the distracting detail that are part of physical world networks.

You are probably already familiar with using special tools to do analysis in your everyday life. Imagine that you are considering buying a new car and are evaluating one that is ten years old. It has 150,000 miles on the odometer. Knowing that 12,000 miles per year is the national average, you want to calculate the average number of miles that were driven on this car. You also know that the transmission may need repair soon. You would like to know the maximum amount you would have to pay to get it fixed. Of course there are many other things that you could do to analyze how well this car would fit into your current budget. The important thing in this example is that you are using statistical concepts like average and maximum.

Since you are familiar with tools of analysis, you will be adding to those a new set of tools and concepts that are related to the area of networks. By analyzing the relationships in a network we can gain insights to take more profitable actions. A few examples will help to understand why this is important.

> A marketing manager in a candy company wishes to get people to try a new candy bar offered by her company. She has some free samples to distribute but rather than give them out randomly, she would like to give them to people who are influential so that if they like it, they will pass the word on to their friends. She plans to use a social network to find the most influential people to receive free samples.
>
> A biologist studying an ecosystem, is concerned the changes in the populations of the organisms that make up the food chain in it. He would like to know the consequences of one particular species dying out. Using computer network models he will evaluate the probability of many different possible scenarios to reach his conclusions.
>
> Investigators at the CIA use network analysis tools to study the activities of suspected terrorists and to identify new suspects. They use phone records and data from other sources to build the networks. The analysis is used together with traditional methods of surveillence and fact checking in order to build cases and stop terrorists before they are able to act.

These are just some examples of the networks that surround us. Most of the time we are unaware of them even though we play an important part of these networks. Like the examples above, though, there are many times that it is helpful to analyze them on a deeper level. The many tasks involved in analyzing networks that we will discuss in the coming chapters include:

- extract the data from the physical network
- encoding the information into an electronic form that can be used with computer programs
- define maintenance activities when building and making changes to the network
- identifying and calculating desired metrics
- run specialized programs to group data, rank objects or do other sophisticated processing

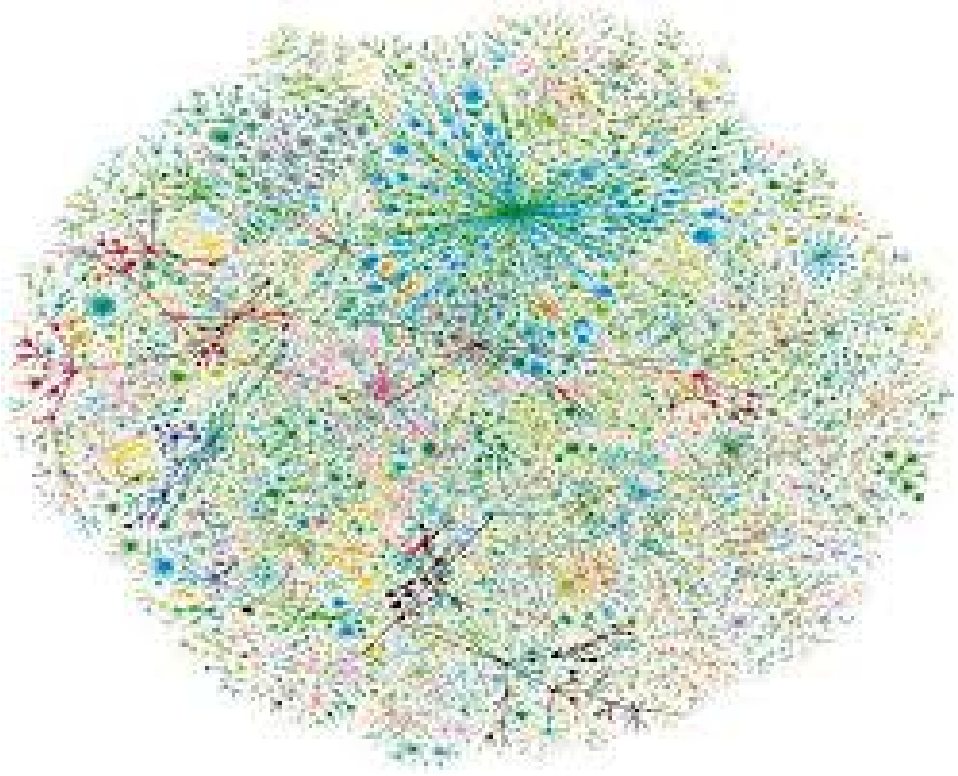

Figure 1.1: Network representation of the Internet

- study the network for growth patterns

In the chapters following we will discuss the processes, metrics, tools and models to learn how to analyze networks. This knowledge will be helpful in a wide range of professions.

1.1 Basic definitions

Network analysis is the culmination of work from many different disciplines, such as sociology, mathematics, statistics, computer science and physics. Because the contributions come from many different areas there are often many names for the same concept which can be confusing. In this section we will define important terms and compare those definitions to others currently in use.

A network is an abstract representation of a real network. Consider a small social network of friends (see Figure 1.2). The objects of the social network (people) are represented in the network as *nodes* (small circles). The lines connecting them are called *links* and in this case we will let them represent the friendship connection between people. You are already familiar with another kind of network, namely a

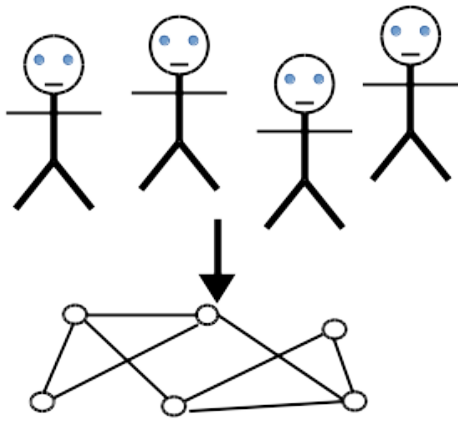

Figure 1.2: Network converted from social network

Table 1.1: Example attributes

id	age	income	grad	years
343	45	2600	y	1
356	26	1300	n	2
387	31	1900	y	3

road map, where the roads are the links and the nodes are cities or intersections of the roads.

Many people refer to the drawing in Figure 1.2 as a graph. For the most part, in this book, we will refer to it as a network but there may be reasons, occasionally, where it is more convenient to use the term graph. They are typically interchangeable but technically a graph represents only nodes and links but that a network can also represent the attributes (defined below) of a node.

Nodes in a network are almost always unique - an object is represented by one and only one node in the network. They can be given a numeric *id* and/or a *name*. The name can be an identifying label such as the person's name. In other disciplines nodes are referred to as points, vertices, or actors.

In addition to the id and name, nodes can contain *attribute* data. This is data associated with the node representing properties of the node. Attribute data can be stored in a table like in Table 1.1. This is typically how we think of data when not considering a network. In Table 1.1, we can see that node (id) 343 is a person who is 45 years old, has an income of 2600, is a graduate and has been with this company for 1 year.

Sometimes we are very interested in one special attribute called the *class*. This is a label we give to nodes that identify them as a member of a particular class. For example, in a social network we could label nodes as "liberal" or "conservative". The class will not be important until Chapter 8.

The links in a network define the relationships between the nodes. They are represented in network drawings as the lines between the nodes. The links determine the structure of a network in the same way as girders and beams determine the structure of a bridge. Without the links, we just have a list of nodes or perhaps a table of the attributes like Table 1.1. Links play a very important part in the network. From a statistical standpoint, it means that the nodes are not independent. For example, from a list of employees, you could select any two and they would not appear to have

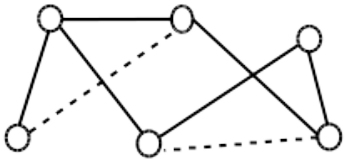

Figure 1.3: Network with 2 different link types

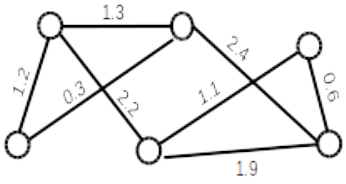

Figure 1.4: Network with different link weights

anything to do with each other. However, if you place these employees on a network with links indicating communication, then it is clear that two linked employees impact each other. Links allow for a rich set of metrics that allow us to measure such things as the distance from one node to another.

In other disciplines links can be referred to as edges, arcs, or ties. They can also have the attributes of type, weight and direction. For most of this book, the networks will be the simplest case, having links of a single type, no weight and directionless. However it is good to keep in mind that other configurations exist and may be important to specific applications.

Networks where links can take on different types are more complex but can capture the data more realistically. A social network with a single type simply indicates that linked nodes have a generic relationship. However, with different link types, relationships like "father of", "mother of" and so on, are easily represented. Many of the algorithms and metrics do not assume different types and so they may need to be modified. In drawings of graphs, types can be expressed using labels, different colors or different line strokes or types (dotted, dashed, etc.).

Having link weights allows a network to represent physical networks that have different capacities for the links. Computer networks, for example, have cables between computers that can carry different amounts (bandwidth) of data. Or a network that represents a road map can use weights for the number of lanes. Again, metrics and algorithms may assume an unweighted network, so changes made be needed.

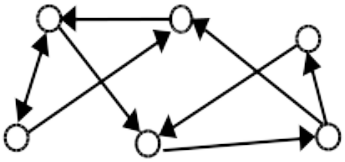

Figure 1.5: Network with directed links

In a directed network, links are usually drawn as arrows to represent a direction in which the relationship flows. For example, in a network of research papers, the papers can be linked by citations, where the links typically are arrows pointed from the citing paper to the cited paper. In graph theory, undirected links are called edges and directed links are called arcs. Notice in the directed network in Figure 1.5, all the links have an arrow on one end except one that has arrows on both. You can consider this two links, one pointing in either direction.

The link properties of weight, type and direction can also be used in combinations. For example, in a water distribution system, the direction of the links can indicate water flow and the weights can be used to show flow capacity.

1.2 Growing awareness of networks

Social networks have existed since, or before, humans became civilized and formed villages. The recognition of social networks, at least in academic literature, surfaced in the early 20th century. It started with a few sociologist interested in relationships in groups of people and eventually grew into the area of social network analysis or *SNA*.

In mathematics, the study of networks started in 1736 when Leonhard Euler published a paper that proved that there was no solution to the Konigsberg bridge problem (how to cross each of the town's seven bridges only once). With this paper, Euler laid the foundation for the mathematical study of graphs, which later came to be called graph theory. Graph theory provides important insights into graphs, often using mathematical proofs.

Not long after the introduction of computers, mathematicians began to formulate graph problems as programs so that these problems could be solved much faster than they were manually. Sociologist also applied computers to their SNA problems. Not only did this mean

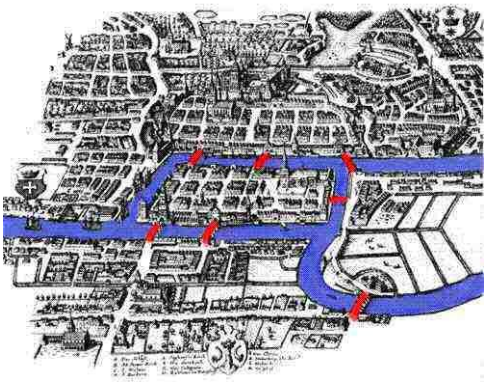

Figure 1.6: Konigsberg bridge problem (source: MacTutor History of Mathematics Archive)

that calculations were faster but it also meant that the networks that could be larger. Prior to computers, 50 nodes was quite large. However, early computers were much slower and had smaller memories so networks were still limited to just one or two hundred nodes.

In the 1990's and early 2000's, several advances were made. First, as Moore's law predicted, computers continued to get smaller, more powerful and faster year after year. Next, the Internet became popular, making it possible to storage large amounts of data on large servers to be shared by people all over the world. Also, the world wide web (*WWW*) was created that brought the idea of a massive network to the consiousness of the average person. Finally, online social networks (mySpace, friendster) became popular giving rise to large scale social media captured in electronic format.

As a result, businesses, researchers and ordinary Internet users became interested in networks. As interest increased, people began to look at data in new ways and the analysis led to new insights. Some examples:

- given a list of movies, the cast of actors and directors for each movie one could convert the data to a network. The actors – or nodes – are connected by movies that they costarred in. Now one can quickly find out which actors worked together on the same movie. Beyond this, using techniques described later in the book, one can even predict which actors maybe likely to work together in the future based on past experience.

- relationships between words could be studied better by creating a network, draw-

ing links between words that are often co-located in the same sentence or paragraph. Using such a network, algorithms could suggest words to users as they type.

- collections of published papers have been converted to networks by using the citations from one paper to another. These networks can then identify papers that probably share a common topic, find communities of authors in the same area, and recommend likely coauthors.

These examples would have been possible before the 1990's but the availability of the data, the ability to easily send it around the world and the power to process such large data sets, made it possible for a single person to do the work fairly quickly. Going forward, it is likely that the interest will continue to grow. Organizations with private data stores also have many reasons to analyze their data as networks for similar reasons. Analysts with the experience and tools to study networks will be on the forefront of exploration in this area. Even having a basic knowledge of the tools and metrics used in network analysis will be a valuable asset.

Chapter 2

Survey of Networks

Networks are woven into nearly every facet of our existence. Our families, friends and colleagues at work are organized into social networks. The products that we sell and buy are built and distributed using networks. Networks are represented in the smallest biological ecosystems and the largest celestial bodies. In this chapter we will take a brief tour of the many networks that are part of the world around us.

2.1 Nature of networks

Before we look deeper into the networks around us, we will get started with a discussion of the nature of networks. Networks are just one of the many ways to represent information. There are also lists, streams, tables and many others. As an example, lets think of the different ways to represent our friends. First, we can think of a simple list of our friends. This could be helpful in deciding who to invite to a party. A table could also be used, like a spreadsheet with a different friend in each row and the columns representing information associated with each friend (like address, phone, and birthday).

Lists and tables are very effective ways of storing and looking up information. The problem is that they make a simple assumption that the data is independent and does not have relationships between them. For example, with a list of friends, two of them may be related by birth and others may be friends with each other. Networks provide a more complex representation that captures the individual relationships between the objects of the network. It is important to keep in mind that when looking at a collection of data there are different ways to organize it. When we choose to organize it like a network it is because we wish to place importance on the relationships between the objects.

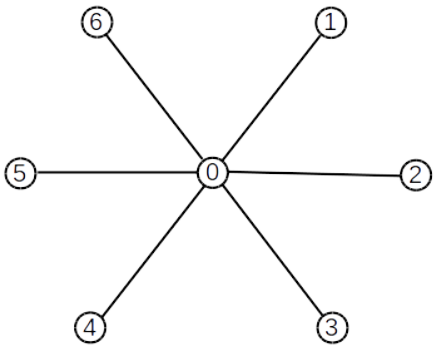

Figure 2.1: Star

As we look at the example networks you will notice that they have different structural properties. For example, some are constructed around stars, where a particular node has several neighboring nodes. Then, in turn, these nodes also have neighboring nodes. Eventually it becomes a network, but mentally we construct it one star at a time. Other networks are called two mode networks because the nodes are one of two types. As an example, we could start with a list of college professors and the students they advise. As we will see later in the book, a one mode network between just the students can be constructed from this two mode network. Of course, there are also networks that we observe directly. For example, it is possible to draw a graph of the people we know as nodes and the links between them representing the friendships. Lastly, there are networks that are built from overlapping cliques. A clique (which we will discuss later in the book) is a subnetwork of nodes, where every node is connected to every other. As an example, one could build a network of a com-

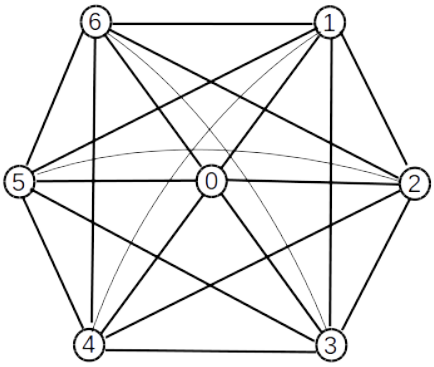

Figure 2.2: Clique

Figure 2.3: Computer network

pany this way. You could assume that everyone in each department all know each other, and thus create cliques for each department. From there, you could connect employees from different departments based on relationships that cross between the departments.

2.2 Example networks

Most of the examples that follow will be familiar to you but perhaps not as formally defined as networks. For each, we will discuss the aspects of the system that are important to a study of networks.

2.2.1 Electronic

Computer networks are a common fixture in educational, industry, commercial and government workplace. Computers, routers, printers and other devices are the nodes that are connected by twisted pair, coaxial or fiber optic cables. Computer networks are built to make it convenient for users to share data and devices and to communicate with each other. Typically the devices communicate with each other, one node to another, however, they do also occasionally send broadcast messages to all nodes on the network. For a given network, the links are usually bi-directional and the same for all connections so we do not think of the links as being weighted or having a direction. Computer networks need to be expanded over time and maintained so it is important to keep diagrams of the network available. Thus, information about the network structure is readily available.

The electronic components that make up networks are prone to failure occasionally. It is important to design and maintain these networks with care so that service can be provided with minimal interruptions. Analysis of the networks helps network engineers and administrators to avoid such problems when computers or link media fail.

Two computer networks that are connected are referred to as an inter-network. In the 1960's the US government commissioned the development of a packet-switched network for communication and protection of sensitive military and academic data. Over time the Arpanet grew by connections to new computers and networks until it came to be referred to as *the* Internet.

In a network, computers can be directly connected; networks are connected to other networks by routers. These machines are simple computers that have connections to two or more computers (belonging to one or more networks). They receive messages on one line and then send them out on another. They have complex algorithms that help them determine which line will be the best for passing along a particular message.

The Internet is not owned by any organization nor is it centrally governed. The various networks that make up the backbone are self governing. It grows by connecting computers that were not part of the Internet to a network or computer that is part of it. As long as new computers and routers support the protocols (rules for sending/receiving information) of the Internet, they should function adequately.

The basic method of sending/receiving information is through packet switching. Large messages, web pages and other documents are divided into smaller *packets* which are sent from one router to another to get from sending to receiver. The packets travel independently through the network. There is no guar-

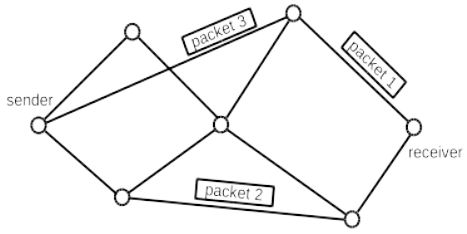

Figure 2.4: Packets moving through the Internet

antee that they will arrive in the right order and some may even get lost (dropped). Part of the protocol stack of the Internet is responsible for keeping track of the packets to make sure they all get to their destination in order.

With networks that are designed and maintained by one organization, it is simple to gather information about the network - it is in fact part of the design and maintenance process. However, the decentralized nature of the Internet has advantages that make growth and maintenance of the network simple but it also makes it difficult to gather statistics or even perform a simple count of devices. Without a central governing authority analysts must resort to using utilities in clever ways to chart the Internet. One utility, traceroute allows one to trace the route that a dummy packet takes from a source node to a destination node. Using traceroute millions of times between many sets of nodes, one can construct a network that approximates the structure of the Internet. Because of the sheer size of the Internet and that many individual computers change IP addresses often, this is normally done just for routers and the main computers of ISPs.

Another example of electronic networks is the network described by the components and their connections on circuit boards of computers and other electronic devices. It is important to the designers to make circuit boards that are inexpensive and high performance. This means placing components close to each other to make the connections as short as possible. Designing them using a network model is helpful and can also help to analyze them after they've been designed.

As mentioned above, it is important to analyze networks to keep them reliable. Analysis is also helpful in making them efficient. Routers maintain tables of destinations that they continually update to help them decide which direction to send packets. It is also helpful to run programs to calculate shortest paths through a network.

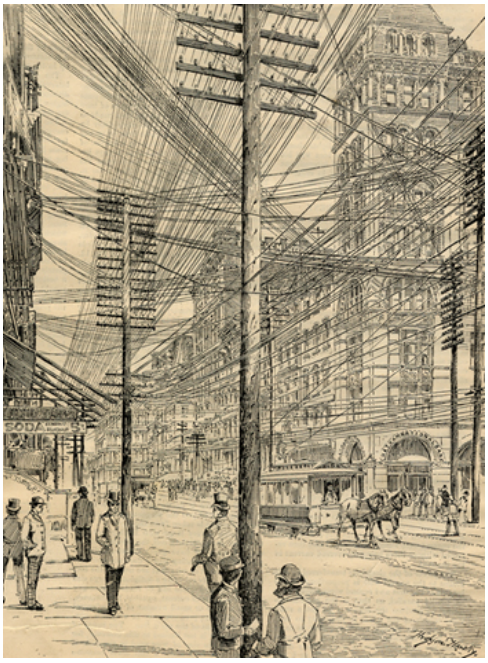

Figure 2.5: Telephone lines

2.2.2 Utilities

Modern society has become accustomed to the conveniences and comforts of the appliances and fixtures that populate our houses. These are not possible without connections to water and sewer lines, the electric grid, gas pipelines and telephone and cable TV wires. Most of these can be modelled using networks. These are operated by municipal or investor owned utilities. These organizations maintain maps and other documents that describe the networks. While these documents are mostly unavailable to the public it is still instructive to study them.

Telephones have been available to the public since the late 1800s, after Alexander Graham Bell demonstrated that sound waves could be converted to electrical signals and sent over wires. Over the years, telephone companies have built massive telephone systems including the high capacity cables between large urban centers, the wires that go to each commercial and private customer, and the switches, computers and software that make the connections between callers.

The model of the wired phone system is simple in theory: two callers are connected by a cir-

cuit and their voices are carried over the wires of the circuit. Of course, while the phone companies could wire each customer to each other customer but that would be a massive, complex and unnecessary system. Modern systems consist of wires from customer premises to a switching station. Cables connect the switching stations so that a customer can call any other customer in the world using the others phone number; the switches then set up the virtual connection that creates the virtual circuit between callers.

The phone system can be analyzed on two different levels. In both levels, nodes represent callers. First, there is the physical level where the links are based on the physical wires and cables that make up the phone systems. The switches themselves would be nodes as well as the callers. Second, the virtual level can represent the virtual circuits created by the phone calls. Given a specific time frame, one can create a network by placing links between two callers if, during that time, the two participated in a phone call.

Note the difference between these two types of networks. In the first, all customers are connected directly to their local switching stations. This is a special structured called a star where the main node is connected to all the satellite nodes and the satellites are connected only to the main node. The switching stations are more like a typical network, where every station may have many connections to many others. In the second (virtual) network, the links between nodes are more typical of a social network. Indeed, if phone calls are indicative of friendship or business relations, then it does in fact, reflect a social network.

Analysis of the phone network can be helpful for planning. While phone companies will often have availability targets of 99.99% or greater, those targets can be achieved because an any given point in time, a small percentage of users are actually on the phone. In times of disaster though, the phone systems become flooded with calls. Having models of computer networks allow companies to create action plans for high call volumes. Another example of constructing networks for analysis from phone records is when information agencies request phone logs from companies to help them track the movements of potential terrorists.

In the past few decades, cellular phones have become popular and now are supplementing and even replacing wired phone service[56]. Cell phone systems are similar to wire system except for the connection to the individual customers. Instead the cell phones make a radio connection to a local cell tower. When a caller makes a call to another, the circuit is connected from from their phone to the cell tower and from there it goes through a wired connection to a switching station like the wired system and the connection travels across the switches until it is connected to the other callers premises or to a local cell tower.

Figure 2.6: Cell Phone

Inferring a network from the physical devices is not really possible for cell networks because the devices are not physically wired. Also problematic, cell phones can be moved from one location to another and be turned off or on. Like wired phones, though, a virtual network can be created based on the calls between callers.

The systems that provide water, sewer and gas are designed in a special form of network called a tree. Water is typically pumped into resevoirs that supply customers with water on demand. Unlike phone systems, where the service is based on being able to connect different customers, connections between customers is not allowed. While two neighbors can be connected to the same water line, the flow of water is strictly from the supply to each customer. Gas companies operate under this same model. While the flow is reversed for sewer systems, other than that it is essentially the same, sewage flows directly from customers to treatment centers.

Water, sewer and gas are primarily transported underground. When a telephone line is broken, it can be found quickly and easily (although it does involve the difficulty of climbing a pole). Finding leaks in water and gas lines

is more difficult though and is one of the many reasons the utility companies maintain current maps of the distribution networks. It is also strategically important to analyze these maps for disaster plans and potential terrorist acts.

While electric companies appear to have the same model as water, sewer and gas utilities, there are differences. Most customers need low-voltage electric service, but it is more efficient to transport electricity between major hubs using high-voltage lines. The low-voltage portion of the network is similar to the other utilities described above. The high-voltage lines, however, form a mesh-like network called the electric grid. One reason for this is because a mesh provides a more robust system if a particular line goes down[64]. Another reason is that service companies are often connected to other service providers nearby so that electricity can be bought and sold for economic reasons.

2.2.3 Travel and Logistics

Roads are built to allow people to travel from one location to another; main roads and highways connect cities and villages. Country and provincial maps give travellers a visual representation of the road system. Converting it to a network can allow a detailed analysis of traffic patterns, finding good routes for a particular journey or simulate the effect of construction on travel times.

Railroads can also easily be converted to networks. The physical layer of cities and rails is a natural way to construct such a network. However, some trains are local (stop in every station on the rail) and others are express (only stop in larger, metropolitan areas). Thus it can be more realistic to connect two stations only if there is a direct train between those stations. In a similar way, airline maps can be recreated from logs of flights, which are freely available. Cities are connected by direct flights between them.

In addition to the utility distribution systems covered earlier, there are also distribution systems that deliver products to locations along the entire supply chain from raw material to end user. These systems use the road, rail, and airline flight networks as their backbone but it is more effective for organizations to use only the locations and routes that apply to their needs and resources. For example, an automobile company could create a network with nodes for their 25 factories, 100 vendors and 250 dealers. Then they would connect them using the railroads and trucking companies that they use to deliver parts and autos. Likely they could weight the links based on distance or tonnage capacity.

Political maps are also easily converted into networks by using the political unit (country, state, province, etc) as nodes and borders as links. For example, in Europe, the node France would be linked to Belgium, Luxembourg, Germany, Switzerland, Italy, Spain, Andorra and Monaco. Such networks could be helpful for planning related to cross border travel and other jurisdictional matters.

2.2.4 Social

Friendship networks have been around for a long time although sociologists only began to study them in the early 20th century. Data for social networks, though, exist in books and records going back hundreds of years. Modern researchers have constructed networks from the works of Charles Dickens (David Copperfield)[66] and Victor Hugo (Les Miserables)[49]. History is no obstacle as long as there is a record. Researchers have been able to piece together a social network for the province of Lot in southwest France using legal documents [13].

Networks have also been constructed from more recent data in direct and indirect ways. Davis, et al. [20] built a two mode network of the relationships between women based on newspaper reports of social get-togethers. The famous karate network was built by Zachary [92], by observing the relationships in a karate studio in the 1970's. A study that was done by Bearman, et al. [8], followed a group of high school students for 18 months and chronicled their dating relationships. Networks constructed from this data are social in nature but have interesting differences from other social networks. This is due to the mainly opposite-sex nature of dating. In most networks, two nodes that are both linked to a common 3rd node are somewhat likely to be linked to each other. This happens much less often in a dating network.

Social networks can also include working relationships. Work networks include relationships like co-worker, supervisor/subordinate, vendor/customer and other similar types. Organizations are sometimes interested in recording their employee networks to track dissemination of information. They can also use the network to identify important individuals in terms

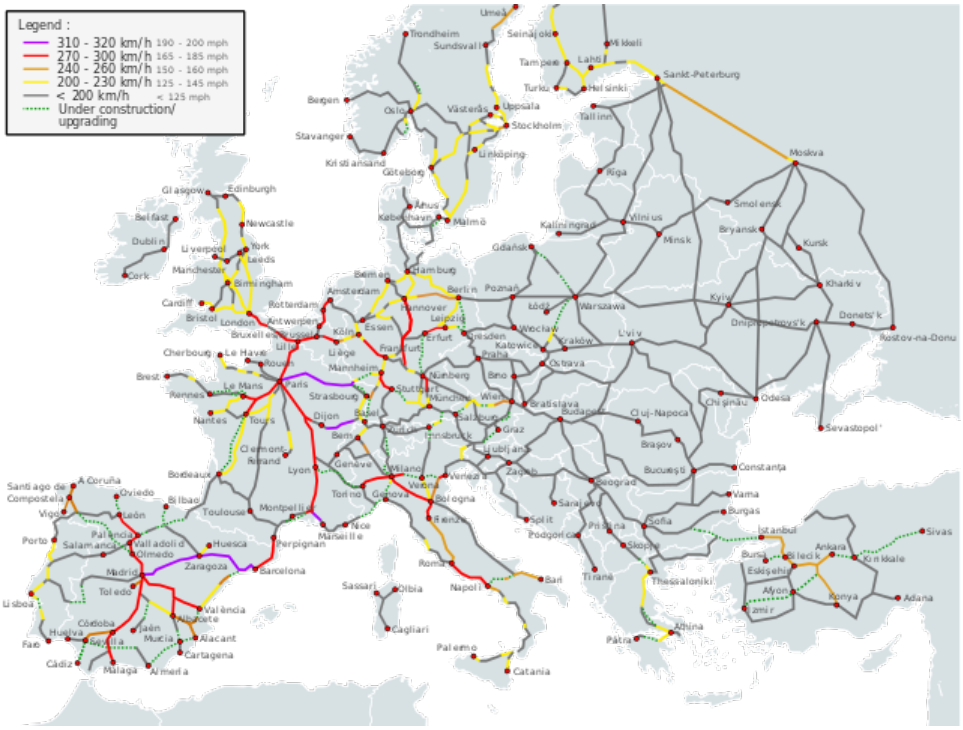

Figure 2.7: European rail network

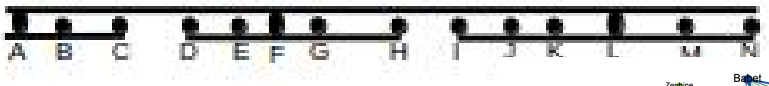

Figure 2.8: Express/local: note that the express train stops at stations A, F and L while the local trains (below) take passengers from the main stations to the smaller, local ones.

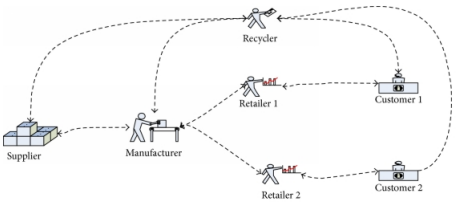

Figure 2.9: Supply chain network

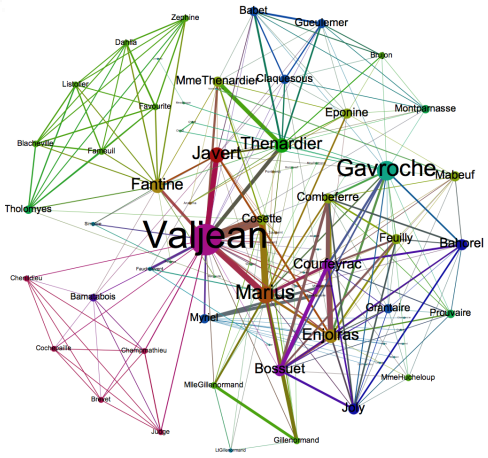

Figure 2.10: Les Miserables

of communication flow.

On-line networks are also available for people to make business connections such as LinkedIn. Many of the contacts people make in these networks are the weak tie type of links discussed in Chapter 6 that are helpful with finding new jobs [26]. Individuals can also use these networks to help find new employees and potential business contacts.

On-line networks with an expressed mission related to shopping, hobbies and other activities have also proliferated. Members of these networks, look for recommendations from other members who have similar tastes and interests. Since these networks are populated by people with interest in purchasing certain products, companies that sell those products are normally interested in having their products somehow represented in the networks. They can attempt to advertise in the network or even offer discounts or free products to highly influential members of the network [72].

2.2.5 Referral

The networks in this section are not ones like in the previous section where people or algorithms make recommendations. The term referral is meant to mean that the actual link is a referral relationship. For example, one can think of a network of referrals from one doctor to another in a particular town. If you ask your family doctor about having surgery to repair your knee, she will probably give one or two referrals for an orthopedic surgeon. One could connect all of the doctors in your town by these referrals into a directed network.

Probably the largest referral network is the World Wide Web (WWW). Web pages are really just files that reside on the hard drive of a web server. The files contain *html* (or some other) code that describes the contents of the page and how it should look. Hyperlinks can be included in the code, which are the links users click on to go from one page to another. The network or graph of the web can be constructed by using pages as nodes and placing directed links from page A to page B if A has a hyperlink pointing to page B.

There is no central controlling agency for the web so no one knows for certain how many web pages there are. One estimate is that there are approximately one billion web sites. Sites can have just one page or several thousand so the number of pages could be as high as a trillion. Finding a page would be nearly impossible without search engines such as Google. The pageRank algorithm that is at the heart of Google's search engine is an example of technique called ranking, discussed in Chapter 8.

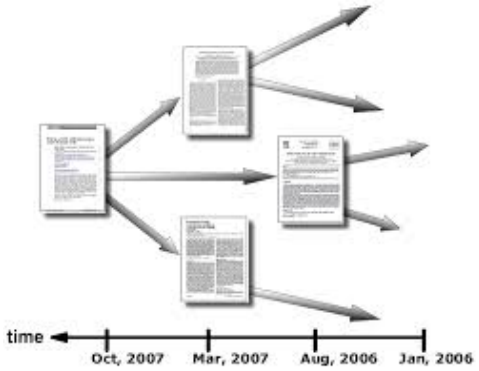

Figure 2.11: Citation network

Citation data sets are large collections of academic papers with citations to other papers. Like the web, one can create a directed network by creating nodes for each paper and directed links from the citing paper to the paper cited. One unusual characteristic about citation networks it the element of time. In the web, often page A will link to page B and page B will link back to page A. The metric called reciprocity measures this property. In citation networks, reciprocity is always zero because if paper B cites paper A, that means paper A has already been published so it cannot cite paper B.

Another kind of referral network also resides on the Internet, like the world wide web. Peer-to-peer (P2P) networks allow users to share digital media like music and movies. With these networks, there is not a central authority or controller. Instead, each member has a piece of software on their computer that has links to a number of neighbors. Those neighbors have links to other members and so on.

2.2.6 Science

Networks are a frequently occurring phenomena in nature. Living organisms have relationships with mates and family members, with other organisms that could be potential predators or prey. Attractive forces such as gravity act on pairs of objects which allow physical systems to be modeled by networks. Science in general makes use of networks for modeling many systems. To illustrate, we offer here a small representative selection.

Living organisms make use of chemical pro-

cesses at the cellular level. The complexity of the interaction of these processes is often modeled using biochemical networks. Enzymes or metabolites are represented by nodes, linked by the physical or functional interaction. By using computational techniques, experiments which were too expensive or impossible to do in the lab are now possible[21].

Ecological networks model the relationships involved in an ecosystem, where nodes typically represent species and are connected to each by the interactions that they experience. Models generally fall into the categories of food webs, host-parasitoid networks and mutualistic networks. The use of computational network models is furthering this area of research [9].

The examples given so far illustrate how knowledge of biological systems can be improved by modeling the system using networks. In a twist, neural networks – the system of neurons, dendrites and axons that allow our brains to control our bodies – are the inspiration behind the artificial neural networks that are used in pattern recognition and machine learning. While the communication in biological neural networks is much slower than in modern computers, complex decisions in humans can be made quickly due to the massive parallelism built into the networks [42].

2.2.7 Data collection

There are many groups of people interested in analyzing networks as well as many different reasons:

- Businesses - there are many reasons for businesses to be interested in network analysis [12]. In marketing their products, they can make use of on-line social networks to position their product optimally. They can also find influential people to encourage to adopt their product. The distribution network that a company uses can be large and complex. Using network analysis, they can look for efficiencies and avoid risk. Analyzing their employee network can help to identify key people who are important to the spread of information. They can also analyze on-line networks and media sites for positive and negative chatter about their company. Finally, companies also use on-line sites to gather information about potential employees and to recruit new employees.

- Government agencies - for good or bad, governments are interested in data collection and mining. The best known examples are in terrorism prevention and law enforcement[1]. Data is collected on phone calls and on-line activity that can be used to create networks of terrorists and criminals. Analysis of these networks is behind the effort to prevent new of acts of terror or to interrupt criminal activity. There are other more mundane networks that are also analyzed like transportation networks. Also, social networks can be used to study the spread of disease[17]. Finally, international and interstate commerce could potentially benefit from networks of political entities and their tax and tariff systems.

Figure 2.12: Google streetview protest - Germany

- Academic - networks have changed higher education in a couple of different ways. First, MOOCs (massively open online courses) have become a new delivery system for education. Students from places and circumstances that would otherwise prevent them from attending a university are now able to take college courses with just an Internet connection. Most of these systems incorporate social networks to help students learn from each other [81]. Universities are also centers for research into the study and analysis of networks.

Information has become a valuable commodity. One can make a point that this is true of the past 500 years but it has become accelerated since the introduction of wide spread computer use and especially with the growth of the Internet. This can be seen by looking at popular web sites that provide content, such as news, weather and sports sites which post paid advertisements. The information that the sites provide draws viewers which then attracts advertisers, who pay for the ads. The payments then cover the costs of operating the site, including paying for the methods of data collection.

In this ecosystem, it is unlikely that one can get data from an organization merely by requesting it. Although that is sometimes possible, like with Wikipedia. They make their information available for download in XML files [90, 54]. However, in most cases organizations tend to keep their valuable data secure.

In this section, we discuss a few ways in which individuals and organizations can acquire data. The most obvious way is that organizations collect the data as part of their daily operations. For example, Amazon.com has a large amount of data on the products they sell and the customers who buy from them. This is true for all manufacturers, retailers, media companies, tech firms and just about every organization. With the availability of more skilled analysts, organizations will start or ramp-up the mining of this data. Analyzing the data will uncover hidden patterns and provide valuable insights for making better business decisions.

Often, though, individuals and organizations are interested in data that they do not have readily available. If the analyst is interested in a particular set of data, and it can be purchased and there are funds available, they can simple purchase the data. Otherwise, they must resort to another way to collect the data. These include, building networks from historical documents, collecting data by surveying or polling people or by using some automated method such as web page scraping.

Using historical data can be done in a number of ways. First, one can comb through dusty old records in the basement of libraries or government buildings and piece together enough information to recreate a network. Another way is to scourer a book, movie, television show or play to rebuild the social network that was central to it. As noted above, this has been done for books by Dickens and Hugo.

Surveys can be done in many ways but face-to-face and telephone surveys are becoming less common. There are many on-line surveys that are available. One designs the survey and tries to interest people in taking it. The information collected then is already in electronic form and easier to convert to a network. There are also services such as Amazon Mechanical Turk [19] that allow you to create a survey or task and then pay the people to complete it.

If the data is freely available from some online source, one can also write their own program to collect the data. A few examples:

- API - some sights offer an applications programming interface (API) that allow programmers to access their data[59]. FaceBook has an API but does not allow for unlimited downloading. On the other hand, Steam, the gaming sight, also provides an API and users can download as much of the network as they want.
- Network utilities - to gather information about the Internet, programmers can make use of the Ping or BGP to discover new nodes and links while crawling the network [64].
- Scraping - some websites do not offer APIs but still do not try to restrict users from gather the information on their site. In this case, web crawlers can be written for web scraping or XML-based extraction [60]. Remember that a web page is (often) a file of html code. The program can be written to download the file and then strip the desired data by looking for specific markers in the data.

It should be noted that because organizations value their data, programmers need to be careful to verify that there are no restrictions on collecting the data. Most sites have a terms of agreement that will specifically state what are acceptable and unacceptable activities.

Chapter 3

Network Representations and Conversions

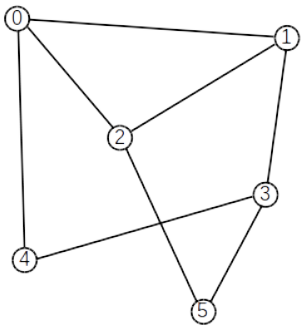

Figure 3.1: Undirected, unweighted network

To help us think about networks, the convention of drawing the network as a graph has become commonplace. While this was a good way to visualize networks prior to the age of powerful personal computers we can analyze networks much more thoroughly and quickly using specially designed software. It is still helpful to visualize networks even if they are large, like the one in Figure 1.1, just to get an idea of the overall shape and structure.

The purpose of this chapter is to discuss the various representations of networks. First of all, they can be drawn as graphs like in Figure 3.1 to the right. This graph could represent a group of six people and the relationships between them (or many other networks of six objects). Graphs help us to manually visualize the network. For example, By looking at a graph of people and their friendships, we can quickly see who is popular and who is not. There are many other simple concepts that become clear to us when a network is converted to a graph drawing.

Table 3.1: Edge List

0	1
0	2
0	4
1	2
1	3
2	5
3	4
3	5

Graph drawings are very helpful to us humans mainly because we can see them. They are not so useful, however for computers, which cannot "see". There are several electronic representations that are useful for computers. Encoding a graph into one of these representations allows us to make use of the many metrics and algorithms that can be programmed into a computer. There are different representations because they each have different advantages and disadvantages. Below we discuss three popular formats for storing a graph. They are drawn as tables but can be stored in programs as arrays, matrices, hash tables or many other appropriate data structures.

3.1 Basic graph representation

To show how graphs can be represented we begin with the simple case of undirected, unweighted graphs. After this we generalize to the weighted and directed graphs.

The first format, is an edge list. As the name suggests, it is simply a list of each edge of the graph. An edge is identified by the two nodes that it connects. For example, the top edge

Table 3.2: Adjacency Matrix

	0	1	2	3	4	5
0	0	1	1	0	1	0
1	1	0	1	1	0	0
2	1	1	0	0	0	1
3	0	1	0	0	1	1
4	1	0	0	1	0	0
5	0	0	1	1	0	0

Table 3.3: Adjacency List

1	2	4
0	2	3
0	1	5
1	4	5
0	3	
2	3	

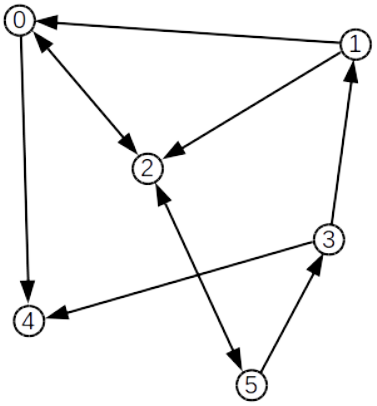

Figure 3.2: Directed network

of the graph in Figure 3.1 that connects node 0 to node 1, is referred to as $e_{0,1}$. The table to the right is one possible edge list that would encode the graph in Figure 3.1. Note that there are eight rows in the list, one for every edge in the graph. Since the edge $e_{0,1}$ is the same as the edge $e_{1,0}$ it is not necessary to include both in the list. Sometimes edge lists do include two entries for each edge – one in each direction – to eliminate any possible confusion. In this case it would include the edge $e_{0,1}$ and the edge $e_{1,0}$.

The second format is an adjacency matrix. This is a two dimensional matrix that has a row for each node and a column for each node. The matrix is filled in with zeros and ones where the ones mean there is an edge and the zeros mean no edge. For example, if there is a 1 in the ith row and jth column, that indicates there is an edge between nodes v_i and v_j. The table to the right shows the adjacency matrix for the graph in Figure 3.1. We often referred to a particular adjacency matrix as "A" and a particular cell as $a_{i,j}$. For our example graph, $a_{0,1}$ is 1 and $a_{2,3}$ is 0. Notice that the diagonal, from $a_{0,0}$ to $a_{5,5}$ is all zeros and that the matrix is symmetric along this diagonal (that is, $a_{i,j}$ is equal to $a_{j,i}$). Undirected graphs are always symmetric like this, so in these cases programs need only work with the upper or lower triangular portion of the matrix.

The last format is called an adjacency list. This list has one line for each node and on each line is a list of its adjacent nodes (or in other words, its neighbors or friends). Figure 3.1, encoded as an adjacency list would look like Table 3.3. Each line represents a specific node, the first line is for node 0, the next is for node 1, and so on. The node itself is not included in the list as a computer can infer it from the line number. So, we can look at the 4th line and see that node v_3 has the neighbors, nodes v_1, v_4 and v_5.

3.2 Directed and weighted networks

A directed graph is one where the links have a direction. Instead of edge $e_{1,2}$ simply connecting node v_1 with node v_2, we say that v_1 points to v_2. In directed graphs, edges are called arcs. Practical examples include the Web where one web page points to another web page and social media like twitter where one person follows another person. For directed networks we can still use edge lists, adjacency matrices and adjacency lists with little change. To illustrate the graph above has been modified to be a directed graph. For edge lists only list the arcs in the direction of the arrow. So $e_{1,0}$ is an arc but not $e_{0,1}$. Adjacency matrices have ones for row where the arc ends. So $a_{0,1}$ is 1 but $a_{1,0}$ is 0. Adjacency lists contain only the neighbors that point to a node.

The edge list is similar to the edge list for undirected graphs except links are listed only once, using the node where the link starts followed by the node that is pointed to.

Notice that by looking at the adjacency matrix we can tell that unlike undirected networks, directed ones do not necessarily have symmetric matrices. If one is given a directed graph but wants an undirected graph it is simple to

from	to
0	2
0	4
1	0
1	2
1	3
2	0
2	5
3	1
3	4
5	2
5	3

Table 3.4: Edge list for directed graph

	0	1	2	3	4	5
0	0	0	0	0	1	0
1	1	0	1	0	0	0
2	0	0	0	0	0	0
3	0	1	0	0	1	0
4	0	0	0	0	0	0
5	0	0	0	1	0	0

Table 3.5: Adjacency matrix for directed graph

convert it. Basically turn all arcs into edges. To do with this an adjacency matrix, one can OR the original matrix with the transpose (if either a_{ij} or a_{ji} are 1, set both to 1).

Weighted graphs have numbers or weights associated with each edge. These can represent distances between cities for road networks or strength of relationship for social networks. These require different modifications to the formats. The graph in Figure 3.3 is another modification of the original graph only with weights added. For edge lists one can simply add another number to the line to represent the weight. It is even simpler to modify adjacency matrices; replace the 1s with the actual weight between the two nodes. However, adjacency lists cannot be modified to work with weighted graphs.

Another possible attribute that edges can

1	2	
3		
0	1	5
5		
0	3	
2		

Table 3.6: Adjacency list for directed graph

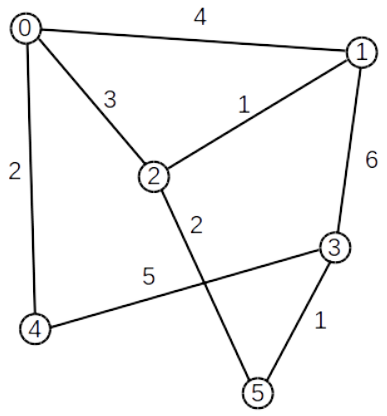

Figure 3.3: Weighted network

have is type. A typical social network will have different types of edges, for example, friend, brother, parent, teacher and so forth. If it is important to encode the type of link into a network, it can be done by assigning a number for each unique type and then encode the numbers in the same way as with weights. It is important to stress that the numbers only represent a category, not a weight. So an edge of type 6 is not 3 times stronger than one of type 2. If such a network is used for analysis using an algorithm designed for weighted graphs, the results will not be meaningful.

	0	**1**	**2**	**3**	**4**	**5**
0	0	4	3	0	2	0
1	4	0	1	6	0	0
2	3	1	0	0	0	2
3	0	6	0	0	5	1
4	2	0	0	5	0	0
5	0	0	2	1	0	0

Table 3.7: Adjacency matrix for weighted graph

Finally, some networks have self links – that is links whose endpoints both point to the same node. For most networks a self link is not meaningful and so it they do not occur often. For example, while there are web pages that hyperlink to themselves this does not add anything meaningful to the graph of the web. Most algorithms also do not consider self links. For these reasons they will not be considered in this book.

3.3 Storage and Efficiency

There are advantages and disadvantages for each of the formats. First consider the storage requirements. Edge lists require 2 numbers (integers) for each edge (or 4 for each edge if both directions are considered). Adjacency lists also need only $2 \cdot m$ integers where m is the number of edges, while adjacency matrices need n^2 integers where n is the number of nodes. To appreciate how this becomes a problem imagine a social network of varying number of members who have an average of 100 friends. See the table below to see how the representations differ:

As the network becomes larger, the adjacency matrix has many more zeros than ones. Most social networks are sparse, meaning that there are few ones compared to the number of zeros. Consider a network like FaceBook: you may have a thousand friends but there are millions of people that you are not friends with. With very large networks like this it becomes more difficult to fit an adjacency matrix into memory.

Another important consideration is efficiency – that is how much time is needed to accomplish specific tasks for the different structures. Edge list are easy to understand and are often used as a format for sending network data between different machines. However, to find out if two nodes are connected it is quite time consuming. Adjacency matrices are perfect for determining edges – to find out if i and j are connected just check the value of a_{ij}. Both edge lists and adjacency matrices are not fast for finding the friends of a particular node. This is what adjacency lists do well. However, they are not as efficient as adjacency matrices at determining if two particular nodes are connected.

3.4 Conversion

Converting from one format to another is a necessary and inevitable task for many reasons. A data set may be given in an edge list but the analysis that is desired requires adjacency list. Sometimes it is efficient to have the data in two different formats. In any case, it is good to have some conversion functions available.

To convert from an edge list or adjacency list to an adjacency matrix, start by creating an n by n matrix with all zeros and then iterate through the list and for each edge i, j, set a_{ij} to 1. If it is undirected one can set both a_{ij} and a_{ji} to 1. Creating an edge list involves appending an edge i, j for each $a_{ij} = 1$ in the adjacency matrix and edge in an adjacency list. The difficulty of creating an adjacency list depends on the data structure used. It the list is an array of linked lists or hash values one simply adds another edge to the linked list or hash. If arrays are used it is a little more work. For each node, it is best to first determine how many neighbors it has, create the array and then set the values of the neighbors. This is not that bad if an adjacency matrix is used. If an edge list is used it might be easier to convert to an adjacency matrix first and then to an adjacency list.

Table 3.8: Edge List

n members	avg nbr. of friends	adjacency list	adjacency matrix
1,000	100	100,000	1,000,000
10,000	100	1,000,000	100,000,000
100,000	100	10,000,000	10,000,000,000

Chapter 4

Characteristics of Special Networks

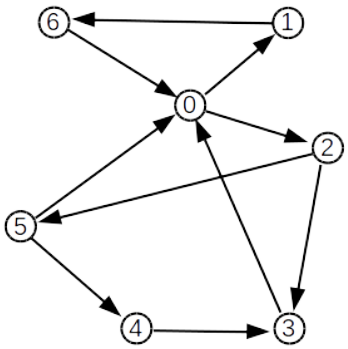

Figure 4.1: A directed network

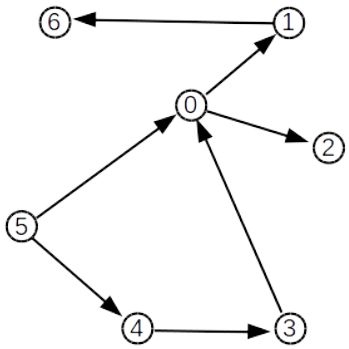

Figure 4.2: A directed acyclic network

When we start to look at networks, one notices that some networks will share similar characteristics. The characteristics of a network influence how we look at it - how we analyse it. Network graphs can be planar or bipartite or some other structure. Knowing the structure will lead us to use different metrics and algorithms to analyse them. A knowledge of network characteristics will also allow us to manipulate the graph so that the analysis is more fruitful. For example, given a directed network, we can calculate many helpful metrics but we cannot use it to find communities (defined in Chapter 8). But we can convert it to an undirected network and then do the community finding. In this chapter, we will explore some special characteristics that some networks have.

4.1 Acyclic directed networks

A directed network is one that uses directed links, or arcs, instead of undirected links (see Figure 4.1). A directed network is acyclic if it does not contain any cycles. In Figure 4.1, the path from v_6 to v_0 to v_1 is an example of a cycle. In some networks it does not make sense to have cycles. For example, imagine a directed network of musicians and their teachers where the links represent "student of". If Beethoven is the student of Hayden and Mozart is the student of Beethoven (he wasn't but assume he was for this example), it would not make sense for Hayden to be the student of Mozart. In this network you would not expect to see cycles.

There is a simple procedure or algorithm to determine if a particular directed graph is acyclic [64]:

1. Search for and select a node with no out-

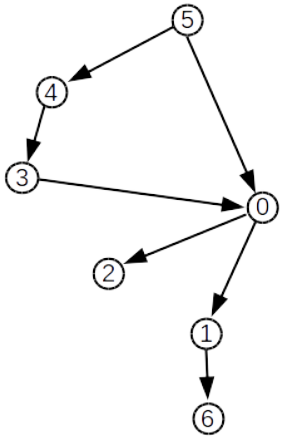

Figure 4.3: Directed acyclic network after placing the nodes using Kahn's algorithm

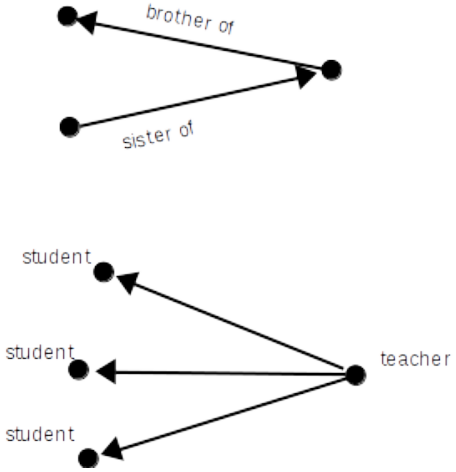

Figure 4.4: Link and node types

going links.

2. If there are none the graph is cyclic. Otherwise, remove the selected node and all of its incoming links.

3. If all vertices have been removed, the network is acyclic, otherwise go back to step 1.

By using the algorithm above, one can redraw the network so that the direction of all of the links all point down. To do this, one simply places the nodes on the paper starting at the bottom with the first selected node and then move up for each selected node. If one thinks of the directed network as having a flow where something is flowing through the network in the direction of the arrows, it is as though the nodes that are at the source of the flow are at top and everything flows down from there.

4.2 Bipartite networks and hypergraphs

As discussed in the Chapter 1, some networks have different types of links. It is also possible for nodes to have different types as well. In Figure 4.4, the top subgraph has links of the types *brother of* and *sister of*, and the subgraph on the bottom has node types that are *students* and *teachers*.

Most of this book assumes that nodes are of a single type or that the type doesn't matter. Examples include the karate set[92], which is a network of 34 members of a martial arts studio and FaceBook. Most of the analysis is done on the graph without considering the type of the node. Most times nodes really do have a type be it is ignored. For example, in the karate set, there are actually 2 instructors and 32 students. In FaceBook, one can assign a type to the node based on many of the properties of the node, such as job title, education level, political designation and so forth.

When we consider a single type for the nodes it is called a one-mode network. A two-mode network (also called an affiliation network) has two different types and there are no edges between nodes of the same type. In graph theory these are referred to as bipartite graphs and have been studied extensively. An example of a two mode network is the Southern Women data set[20]. The data for this set was collected from newspaper stories of meetings of a social club in Natchez, Mississippi. The nodes are of two types: events (meetings) and people. A link connects a woman to an event if she attended that event. There are no links between two women or between two events.

Bipartite graphs can also be represented with a hypergraph (see Figure 4.6). This simply represents the graph as nodes being part of groups. Note that the hypergraph in Figure 4.6 is the same as the bipartite graph in Figure 4.5. The group that contains nodes v_1 and v_4 is actually node v_a in the bipartite graph. Sometimes it is convenient to think of data as a hypergraph first and then switch to the more computationally convenient bipartite graphs. Most people would probably envision the Southern Women data set as a hypergraph first.

It is possible to convert a two-mode or bipar-

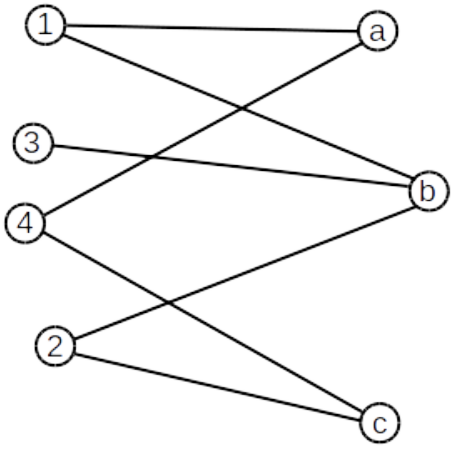

Figure 4.5: Bipartite network

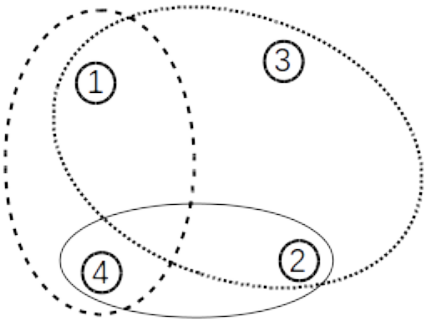

Figure 4.6: Hypergraph network

tite graph into a one-mode graph easily. In converting, the network keeps one of the sets (or types) of nodes and discards the other. Note that the links are necessarily removed. Then, new links connect nodes that are linked to a common node in the bipartite graph. For example, in Figure 4.5, we can remove nodes a, b and c and then draw a link between 1 and 2 because they are linked by b.

In the same way, women in the Southern Women set can be linked together if they both attended at least one meeting together. The rationale differs from one set to another depending on the nature of the nodes and the links. In this case, it makes some sense, because two women attending the same (small) meeting would probably have some sort of relationship with each other.

Many of the real networks are one-mode networks where the nodes are of the same type. Social networks, computer networks and food chain nets are good examples. There are also many two-mode networks. Examples include: a network of movies and the actors that starred in them, researchers and the papers they coauthored, students and the professors they have taken and animals and the habitats they share.

4.3 Trees

Trees are special kinds of graphs that have no closed loops. Most of this book concerns graphs that have loops but there are many interesting features of trees that are worth discussing. For example, when considering graph traversal, there is a problem with "going in circles", that is following a path that leads back to nodes that have already been discovered. This is not a problem with trees since there are no loops. There are many naturally occurring trees found in nature and modern civilization, such as rivers and utilities.

When drawing or observing a tree, often there is a node at the top that is considered the root node. This is arbitrary, as any node can be chosen to be the root node. For example, using the tree from Figure 4.7, one could select the node v_5 (or any other node for that matter), drag it to the top and allow the other nodes to "hang from it".

There are times when it is helpful to convert a graph into a tree by removing links. For example, in a large social network of friends, it may be helpful to create a call-tree. This is useful when it is important to quickly spread

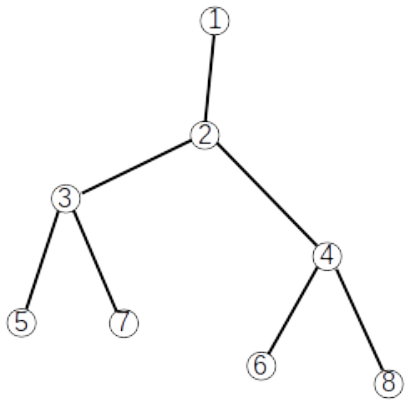

Figure 4.7: Tree

news throughout the network. One person is designated to call another 2 people, those two are each designated to call two others and so forth.

A special method of creating a tree from a weighted graph will result in a *minimum spanning tree*, which connects all nodes with the minimum weighted total of all links. An MST can be helpful for grouping nodes by removing links to separate the nodes into components[52].

4.4 Planar networks

Planar networks are those where no links cross each other. This is not a property that is important in the analysis of networks because it is a visual property and so not a hindrance to software programs. However it is a property that occurs in may real networks.

Planar networks are familiar to us in the real world. The road network is mostly planar, although one can argue that there are roads that cross each other (using bridges) and do not intersect. Geographic areas, such as countries or states, can be mapped to networks. For example, in the US, each state can be a node and states that border on each other would be connected by a link. These are necessarily planar graphs.

Chapter 5

Metrics

Analysis by its very nature requires that a particular process or object be disassembled and scrutinized. Good analysis is supported by tools and metrics. Metrics are the subject of this chapter. They are objective measurements of the subject matter. Like many other fields, in the analysis of networks, the metrics draw heavily from the realm of statistics.

With lists or tables of data, analysts use many familiar metrics such as mean (average) and standard deviation. Those metrics do not apply due to the complex structure of networks. The rest of this chapter presents some of the more popular metrics used in analysis of networks. The metrics are divided into sections based on the object of their application: node, node-pair, or network.

5.1 Node Metrics

Typically, nodes are the first class citizens of networks. They are the people in social networks, the authors in bibliographic networks and animals in the food chains. Links are important but are often not the subject of the analysis. The metrics that follow, tell analysts how centrally located a node is or how authoritative or influential it is.

5.1.1 Degree

Degree is the term used in Graph Theory as a count of a node's neighbors. For undirected networks it is simply the number of links connected to a node. For directed networks, there is both inDegree (the number of incoming links) and outDegree (the number of outgoing links).

In a social network, the degree is the number of friends that a person has. This by itself, reflects the popularity of a node. An analyst looking for an influential or authoritative node can use the degree as a fast and simple measure.

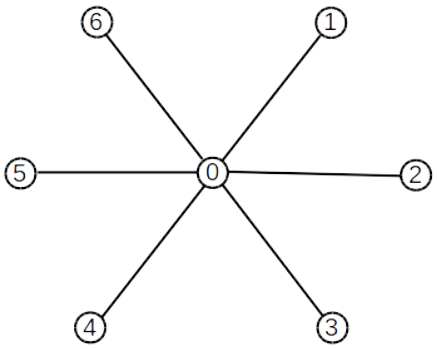

Figure 5.1: Star

5.1.2 Closeness

Closeness is another measure of centrality like degree. The definition is a little more complicated though. It uses the concept of geodesic path between two nodes which is defined under the node-pair metrics. It basically means the shortest distance between two nodes.

Farness is defined as the average of all the geodesic distances between it and every other node in the network. Closeness is the recipricol of farness. Mathematically it is defined as:

$$C_c(i) = \frac{n}{\sum_{j \neq i} d(i,j)}$$

where $d(i,j)$ is the length of the geodesic (defined under Section 5.2) path between v_i and v_j. Note that $0 \leq C_c \leq 1$.

Generally, the idea is that the lower a node's closeness score, the more centrally located in the network it is. Nodes on the fringe of the network will have a high score because their geodesic paths will have to go through the entire network to get to the ones on the other side. In the star network in Figure 5.1, node v_0, in

the middle will have a closeness of 1, whereas all of the other nodes will have a closeness score of 6/11.

5.1.3 Betweenness

Another measure of centrality is betweenness. Betweenness measures how many geodesic paths pass through a particular node. This sounds like the closeness definition but it different. Measuring the closeness for a node (v_i), we are only concerned with then $n-1$ geodesic paths from v_i to every other node in the network. For betweenness, we are interested in how many of all of the $n(n-1)/2$ geodesic paths between all node pairs, pass through v_i.

This presents a dilemma. Sometimes there may be more than one geodesic path between two nodes. Consider two nodes v_j and v_k in an undirected network. Assume that the shortest path between them is x links long and that there are actually three such paths. What if one of the paths passes through v_i and the other two do not? In this this case, the value of $\frac{1}{3}$ is added to v_i's score.

The mathematical definition is

$$C_b(i) = \sum_{jk} \frac{n_{jk}^i}{g_{jk}}$$

where g_{jk} is the number of geodesic paths from v_j to v_k and n_{jk}^i is the number of geodesic paths from v_j to v_k that pass through v_i.

5.1.4 Unconnected networks

Some networks are not connected, that is, there are some node pairs for which is not possible to connect via a path. The parts of the network that are connected are called components. Unconnected networks have implications for closeness and betweenness, however degree is unaffected. In these networks, the closeness and betweenness scores reflect the centrality of a node within the particular component it is part of. This can be misleading as a node that is part of a small component can have much higher closeness and betweenness scores than one in a large component simply due to being in a smaller component. In these cases it is best to analyze the components separately especially when using metrics that are sensitive to the size of the components.

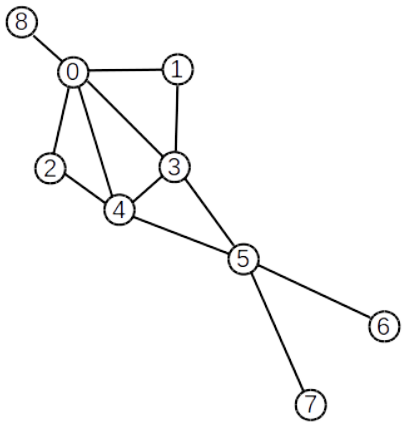

Figure 5.2: Example subgraph to illustrate centrality metrics

Table 5.1: Metric Comparison

	degree	closeness	betweenness
0	5	0.62	0.47
1	2	0.50	0.22
2	2	0.50	0.22
3	4	0.67	0.40
4	4	0.67	0.40
5	4	0.62	0.58
6	1	0.40	0.22
7	1	0.40	0.22
8	1	0.40	0.22

5.1.5 Comparison of Centrality Metrics

Recall from statistics that there are three popular measures of average, mean, median and mode. Just as those three represent a sort of average of the data but differ in important ways, the degree, closeness and betweenness of a node also are similar measures but have important differences.

The three metrics are compared in Figure 5.2. The values for the three metrics are calculated and presented in Table 5.1. Notice that according to the table, the nodes with the highest degree, lowest closeness and highest betweenness score are all different. Node v_0 has the highest degree, nodes v_3 and v_4 have the lowest closeness score and the node with the highest betweenness score is node v_5.

Which metric you use depends upon your interest. In a social network, the high degree nodes have many friends, so we might assume that they would be the best informed and per-

haps good candidates for disseminating information. They might also be most at risk for disease spread. Those with low closeness scores are more centrally located in the network in terms of shortest paths so we might conclude that they have a more balanced opinion of the information circulation in the network. Those with a high betweenness score are often located at important bottlenecks of the network where communities overlap. These nodes can be important for identifying communities and passing information between communities.

Another consideration that is important with large networks is complexity. In computer science, complexity is a relative measure of how much time it takes to compute a particular algorithm. We will not discuss the details of how complexity is calculated but it is enough to know that algorithms with lower complexity will be faster than ones with higher complexity. It is not surprising that degree is the simplest of the centrality metrics and is almost trivial. Next comes closeness because lengths of the shortest paths must be computed. Finally, betweenness has the greatest complexity because it has to actually find all of the geodesic paths, which is much more time-consuming than just finding their lengths. In larger networks the time spent calculating betweenness may become prohibitive.

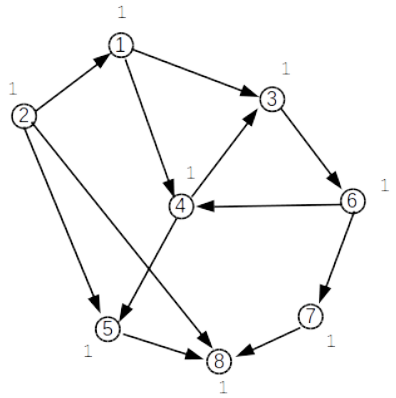

Figure 5.3: Power method: start all nodes with same value.

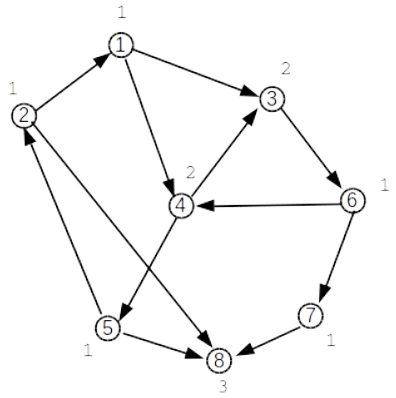

Figure 5.4: Power method: update nodes with sum of neighbors

5.1.6 Eigenvector

There are a number of metrics that fall under the category of eigenvector centrality. The German word *eigen* means characteristic, so an eigenvector is a characteristic vector for a given square matrix. In particular, a vector x is an eigenvector of the matrix A if it satisfies the following equation:

$$\alpha x = Ax$$

The scalar value α is called the eigenvalue. For a given matrix A there can be many pairs of eigenvectors and eigenvalues that satisfy the eigen equation above. Note that while the zero vector would satisfy the equation, it is meaningless and so not considered an eigenvector. The definition and applications of eigenvectors belong to a discipline called linear algebra. The reader is invited to explore it by reading some of the excellent tutorials available on the web (search for *linear algebra tutorial*). However, as this is an advanced, time-consuming subject, it will not be dealt with here. For our intentions, it is sufficient to know that eigenvectors can be used in a wide range of engineering, computer science and general science applications.

As a metric for networks eigenvectors are used to rank nodes according to the importance of their neighbors. It is helpful to compare it to degree. Degree simply ranks nodes by how many neighbors they have. So two nodes that both have a degree of 100 are considered equal in degree. In the real world, though, sometimes we rank based not only on the number of neighbors but also the quality of the neighbors.

Imagine a network of doctors who are connected by links of recommendations. For example, if doctor A and doctor B both recommend each other as qualified physicians, they would be joined by a link. Degree is one measure a how qualified a doctor is. However, if two doc-

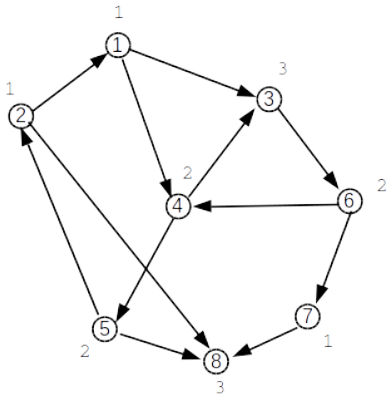

Figure 5.5: Power method: continue to update nodes

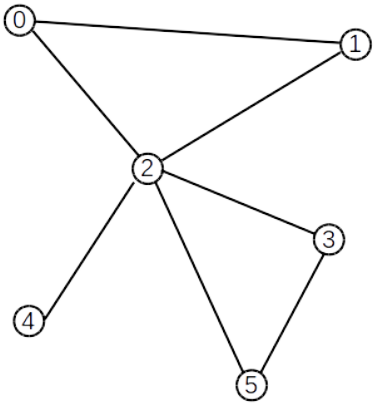

Figure 5.6: Clustering coefficient for the neighborhood of node v_2

tors, X and Y both have a degree of 10, it might be that X's recommending doctors have higher qualifications than those of doctor Y. In this case, even though the degrees are the same, we would say that doctor X is more qualified than doctor Y.

The way the eigenvector metrics work is that they consider not only the neighbors of a node but also the metric values of those neighbors. Figures 5.3, 5.4 and 5.5 show the process for calculating the eigenvector metric. Begin by assigning all nodes the same value. Then go through several iterations, summing the scores of the neighbors in each iteration. The bottom graph shows the result after the 2nd iteration. The process ends when the values converge – that is, they change very little from one iteration to another.

This process is called the power method and is often used to calculate eigenvectors. Of course this process will never end for most graphs because the values for nodes that are part of cycles will just continue to grow indefinitely. One solution is to normalize the values after each iteration. One way to do this is to divide each value val_i by the largest value which forces all of them into the range $0 \leq val_i \leq 1$.

Another potential problem is nodes that have no incoming links. These nodes will go to zero which can potentially propagate through the network. A solution to this problem is to add a small amount to each node when normalizing[69].

5.1.7 Clustering Coefficient

The clustering coefficient measures how tightly connected a nodes neighbors are. At the two extremes are stars and cliques. A star has neighbors that are not connected to each other at all, whereas if all of a node's neighbors form a clique they are all connected to each other.

The mathematical formula for the clustering coefficient is

$$CC_i = \frac{2 \cdot |e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)}$$

where N_i is the set of v_i's neighbors and k_i is the number of neighbors for v_i. The intuitive description of clustering coeficient is the number of links in its ego-network divided by the total possible links. In Figure 5.6, shows the subnetwork inscribed by node 2 and its neighbors (sometimes called the ego-network or ego-neighborhood). The clustering coefficient for node v_2 is $\frac{2}{10}$ because there are 2 links in the subnetwork and there are 10 links possible $(5 \times 4 \div 2)$.

5.1.8 rawComm

Later in the chapter on techniques, we will discuss creating communities of nodes in networks. Generally, these are groups of nodes where the nodes inside the groups are more tightly connected and the links between the groups are fewer. Communities are a way that analysts can describe a network and also to look for natural groupings - for example, within a network of college students, communities may identify

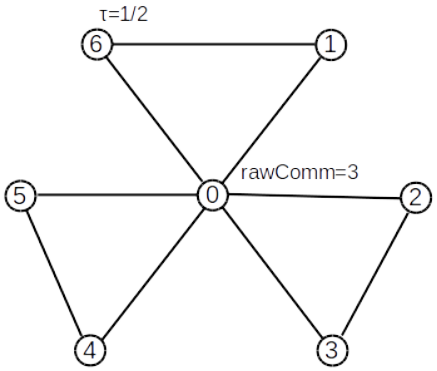

Figure 5.7: rawComm calculation

students with common interests or same course schedules.

There are many algorithms to actually find the communities. Two problems with them is that there is no agreement on what constitutes a good community so there are many ways to find them and they can also be computational expensive to calculate. The metric rawComm provides a way to approximate the number of communities to which a nodes belongs without actually doing a community finding process.

The actual calculation of rawComm is complicated but here we will provide a high level view. Figure 5.7 shows a small, simple network of 7 nodes. To calculate the rawComm value for node v_0, the τ value for each of its neighbors needs to be computed first. In this network, $\tau = \frac{1}{2}$ for nodes v_1 through v_6. Summing all of the τ values yields the rawComm value of 3 in this case. If all of the nodes in this network were connected (clique), the rawComm value would be 1.

5.2 Node-pair Metrics

Node-pair metrics measure properties between two selected nodes in a network. These are often helpful to algorithms to predict new links and finding communities. They can also be used when analyzing a particular node to understand the nature of its connections.

5.2.1 Path length

A path between two nodes v_i and v_j, consists of traversing from v_i to other nodes by following links until reaching v_j. Each link traversed is call a hop. There is a 1-hop path between a node and all of its neighbors (those nodes that it is directly linked to). The neighbors of these nodes (except the originating node) would all be 2 hops away and so on. Some networks are connected, meaning there is a path between any two nodes in the network. If a network is disconnected, then it is made up of components, where all the nodes are connected within the components but there are no paths between nodes from different components.

Between any two nodes, v_i and v_j, there can be many different paths. The paths can be measured either by 1) counting the hops in an unweighted network, or, 2) adding up the weights of the links that form the path. Obviously some paths are longer than others. The term *geodesic path* refers to the shortest path between two nodes. There can be more than one geodesic path between two nodes. Sometimes it is important to know the routes of the shortest paths and sometimes all we are interested in is the length.

When one is interested in the nature of the relationship of two nodes, the geodesic path length can be used to estimate different facets of that relationship. In a social network, for example, it can be used to estimate how likely it is that two people will become acquainted, or how difficult it will be for one of them to get an introduction to the other.

In a famous study by Stanley Milgram[58], a group of people in the midwest were sent envelopes with a name and address written on it of a person in Massachusetts. They were instructed to try to get it to the addressee by handing it to an acquaintance who would then pass it on to another acquaintance and so forth. Along the way, each person would record their name and address on the envelope. Some envelopes got to the addressee and some did not. Of those that arrived, Milgram calculated the average number of people that it took to reach the destination, which was around six, which is the source of the oft-quoted phrase, *six degrees of separation.*

5.2.2 Cocitation, bibliographic coupling and reciprocity

In directed networks, there are two metrics that can be used to measure the relationship between two nodes based on their neighbors. They are particularly helpful in citation networks where links represent one node citing another (for example one paper citing another).

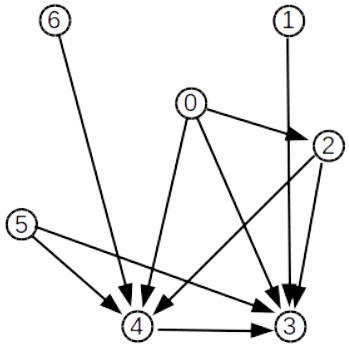

Figure 5.8: Nodes v_3 and v_4 are cited by 3 other nodes

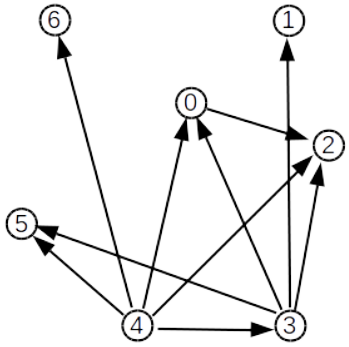

Figure 5.9: Nodes v_3 and v_4 both cite 3 other nodes

Cocitation measures the number of nodes that both cite the same two nodes. Figure 5.8 illustrates this metric, showing that the two nodes v_3 and v_4 have the same citations from 3 other nodes. In a network of publications this is an indication that two papers have similar topics if a large number of papers cite them both.

The metric *bibliographic coupling* applies the same principal as cocitation but using oppositely directed links. Figure 5.9 shows an example of this metric. Again, in a network of published papers, two papers that both cite many of the same papers can also be considered to be about similar objects. A distinction is drawn between the two metrics though. In the case of citation networks, cocitation carries more weight than bibliograph coupling, because being cited is an indication of authority. A paper with many citations to it is considered to have more authority than one that has fewer.

An interesting note is that both of these metrics can be easily computed using matrix multiplication. Given an adjacency matrix A for a directed network, cocitation is the result of AA^T and bibliographic coupling is A^TA.

One last metric that is sometimes used in directed networks is *reciprocity*. This metric measures how many links in a network are reciprocated. If there is a link from node v_i to v_j, we say that it is reciprocated if there is also a link back from v_j to v_i. The actual calculation is the number of node pairs with links in both directions divided by the total number of node pairs with any link. This might be helpful to an analyst studying a friendship network to see if friendship is often mutual.

5.2.3 Common neighbors

As noted above, cocitation and bibliographic coupling apply only to networks of directed links. For undirected networks the corresponding metric is *common neighbors*. For a pair of nodes, v_i and v_j, this is simply a count of the nodes that are linked to both v_i and v_j. Like geodesic path length, common neighbors tells us something about the strength of the attraction between two nodes. The more common neighbors that two nodes have, the more likely we would be to expect them to become linked in the future. This is an important statistic for social networks. Often, with on-line social networks, when one clicks on a member they don't know, the network will display the friends they have in common.

5.2.4 Jaccard

While common neighbors is a well understood and helpful metric it can be somewhat misleading. The problem is that links with some nodes are more significant than others. In our real life social networks, some people may find it advantageous to be friends with well connected people (those with many links), that says something about the links. If you are friends with someone who has 500 friends, we might say that the strength of your friendship may be less than someone else you are friends with who only has 10 friends.

The *Jaccard* metric is designed to mitigate the effects of nodes with differing degrees. The idea is that it normalizes common neighbors by

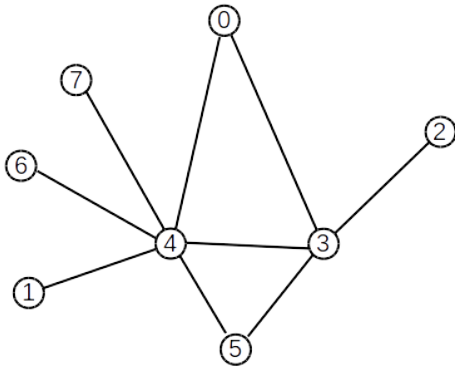

Figure 5.10: Jaccard distance between nodes v_3 and v_4 is 2/8

dividing by the total number of unique neighbors that both nodes have. The formula is:

$$f(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

where $N(v_i)$ is the set of neighbors that are linked to v_i. We will demonstrate the Jaccard metric for the nodes v_3 and v_4 in Figure 5.10. $N(v_3) = \{0, 2, 4, 5\}$ and $N(v_4) = \{0, 1, 3, 5, 6, 7\}$. The intersection of them is $N(v_i) \cap N(v_j) = \{0, 5\}$ and the union is $N(v_i) \cup N(v_j) = \{0, 1, 2, 3, 4, 5, 6, 7\}$, so the Jaccard metric is 2/8. Jaccard is used like common neighbors to predict new links, determine strength of relationships and also for community finding.

5.3 Network Metrics

The metrics in this section apply to networks in general. Each of these statistics measure something specific about a network. There are many reasons for calculating them. Some of the metrics give us a clue as to the overall structure of the network. For example, knowing the number of open and closed triangles provide some insight about the type of communities we can expect to find. As we will find out later, other network metrics can help us to determine the growth model of the network.

Another reason for network metrics is to compare one network to another. For example, say that you are given two different designs for a computer network. If you find that the diameter of one is significantly shorter than the other it means that the one with the longer diameter will potentially take longer to pass messages across the network.

Finally, it is sometimes helpful to periodically calculate some of these metrics on a network that is changing over time. This can be helpful if there are actions that can be taken to provide better service.

5.3.1 Composites

Many of the node and node pair metrics can say something descriptive about the network as a whole by averaging them. The average degree $\bar{d} = \frac{\sum d_i}{n}$, can reveal the nature of nodes in a network. In FaceBook, it is 338 (the median is 200)[71]. The average number of followers per user on Twitter is 208 [23]. The average number of instagram followers is 843 (based on a random sample of 21,239 users)[41]. As a comparison, the Dunbar number [25] (named after anthropologist Robin Dunbar), which correlates the size of a primate's social group to the size of their brain, assigns the number 150 to humans. While these numbers can lead to interesting, late night discussions on the nature of online relationships, clearly the nature of online friends and followers is different than those in the physical world.

In some circumstances an analyst may be interested in a particular characteristic of a network that is best measured using an average. For example, assume that in a particular distribution network, the goal is to reduce the time required to overall delivery of goods. Also assumed that deliveries are made by rail. Obviously the delivery times can be reduced by adding more links (rail lines) but it is expensive to do. In this case, average closeness could be calculated for different configurations of new links. Then the best decision for reducing average delivery time could be the configuration that minimizes the average closeness.

Clustering coefficient tells us how connected a node and its neighbors are. The average can be used as an indication of how well the network can be clustered. To illustrate, imagine two different social networks, one that consists of the students from a modern university and another from a middle ages european country such as France or England. In typical middle ages society, most people were peasants who lived their lives on a single manor. A recent study shows that the small world effect was not present in the middle ages [57]. This means that the short paths between any two individuals that we ex-

perience now did not exist then. For most people, their links would be formed between the friends and relatives on the manor. Most manors were small enough that nearly everyone would know everyone else. Universities do have some small pockets of densely connected students, such as those students that are part of a sports team, club or academic major. However there is a lot of mobility where students take courses outside their major and participate in clubs and sports away from their friends from other groups. The average clustering coefficient will be much larger in the middle ages network (where groups are more clearly defined) compared to the university net.

Recall that geodesic path length is the length of the shortest path between two nodes. Averaging it over every combination of node pairs is a measure of the mobility of the network. Mobility can refer to different things in different networks, e.g., products in a distribution network or information in a social network. Using the prior example of middle ages / university network, clearly mobility would be much higher in the university network. Calculating average path link for both networks would confirm this supposition. The concepts of average clustering coefficient and average path length will be used in Chapter 7 where they are used to define a particular way in which networks can form.

Besides using averages, it is often helpful to use maximums. The maximum degree in a social network tells us who is most popular: how many friends the person with the most friends has. It is also helpful in looking for cliques. Finding cliques is computationally expensive. Finding the upper bound of the degree (d_{max}) of the maximum degree node, though is simple since we know that the largest clique cannot have more nodes than $d_{max} + 1$.

The maximum geodesic path length (over all possible node pairs), has a special name: diameter. The diameter tells us how far it is, in the worst case scenario, that a message has to travel across a network. In 1999, the WWW was estimated to have a diameter of 19 [2]. Since it has been shown that networks tend to have shrinking diameters[51], it is probably shorter than that now.

5.3.2 Density

The density of a network is the number of link-ends divided by the total number of node pairs:

$$density = \frac{2m}{n(n-1)} = \frac{\sum d_i}{n(n-1)} = \frac{\sum d_i}{n} \cdot \frac{1}{n-1}$$

Since each link has ends in two different nodes, the numerator must be the number of links times 2. The last expression is average degree times $\frac{1}{n-1}$, meaning that density is a function of both average degree and the size of the network. To see the significance of this, imagine an on-line social network of 200 nodes that grows over time to have $200,000$ nodes but the average degree stays constant, say 25. The density of the smaller network will be 0.126 while for the larger network it will be 0.000125 [64].

This leads to the definitions of sparse and dense networks. As a network grows, if the density remains constant it is considered dense. If the density shrinks, it is considered sparse. Thus, social networks, in which people have a soft upper limit on the number of friends they can maintain, are necessarily sparse.

5.3.3 Degree distribution and power law coefficient

Average degree and maximum degree are both important statistics as described above. However, they are not that helpful to getting an understanding of the overall structure of the network. To better understand the structure, we can look at the degree distribution. This is simply a list of the degrees of each node, normally sorted in descending order. As an example, consider two networks. The first represents the games played by division one American college football teams, where each team is a node and the links represent the games connecting the two teams that played in the game. The second network is a simple social network of about 100 people. Some people in the network have just one or two friends and some have many more. The average and maximum degree for the football network are both 12 since teams in that division play exactly 12 games. Lets say the average for the social network is also 12.

Even though these two networks have the same average degree the actual structures are quite different. Examining the degree distributions would allow someone to more clearly see the differences. It is often helpful to plot a histogram or a curve to see the distribution easier. In many social networks, the degree distribution behaves according to a power law. An example of a power law curve can be seen in Figure 5.11. The degree is plotted on the vertical axis for each node on the horizontal axis. Note that the nodes must be arranged in descending order of degree size. The formula for a power law curve is

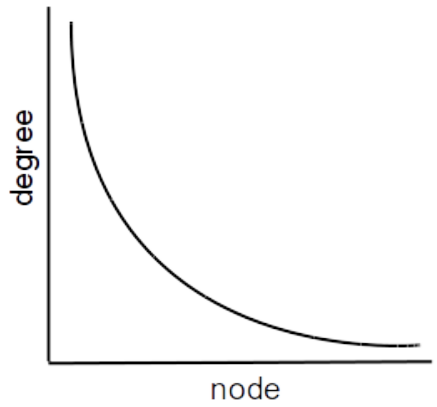

Figure 5.11: Power curve

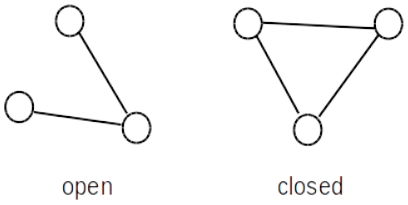

Figure 5.12: Triads

$$p_k = \frac{C}{k^\alpha}$$

where p_k is the probability of a node having k neighbors, C is a constant and α is an exponent, which for a power law should be between 2 and 3.

5.3.4 Cliques

A *clique* is a graph in which every node is connected to every other node. Given a network, we can identify subgraphs based on a subset of nodes and include all of the links that have both ends connected to the nodes in the subset. For a variety of reasons, it is interesting to find subgraphs that are cliques in a network. This is computationally hard for large cliques, so often it is done only for cliques of size 2 through 5.

Groups of nodes of size 3 are called triangles or triads. In undirected graphs, triads can be open - where one node is connected to the other two, which themselves are not connected - or closed, where they form a clique. One can think of this as a limited version of clustering coefficient. In fact, one can define average clustering coefficient as

$$CC_i = \frac{closedtriads * 3}{alltriads}$$

While triads do not give us much more information than clustering coefficient in undirected graphs, they are much more important in directed graphs. Instead of just open and closed triads, in directed graphs there are 16 different types.

Chapter 6

Social considerations

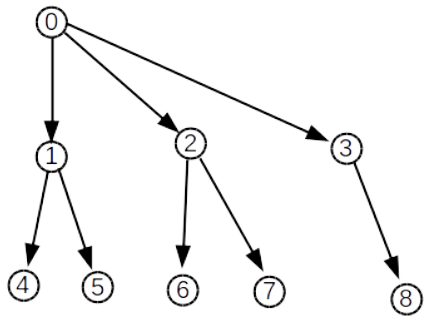

Figure 6.1: Positions of different nodes

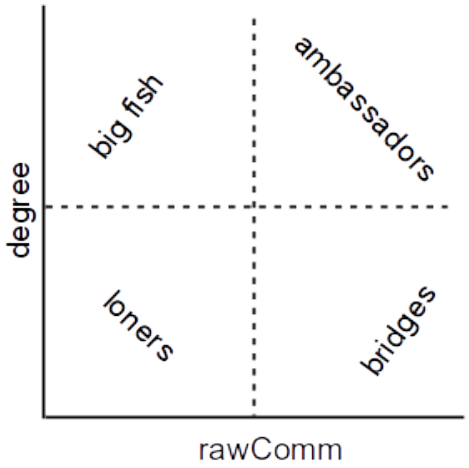

Figure 6.2: Roles defined by rawComm and degree

6.1 Position and roles

Nodes within networks can be assigned various values. Besides attributes, we can also calculate metrics (see Chapter 5). The notions of position and role – while sometimes related to metrics – are two other ways of describing nodes. Without considering networks a person's position is based on their family, education, career and friendships. Their role in society has to do with how they interact with their contacts. Imagine a group of friends where the informal role of one is to keep the group connected while the role of another is to introduce new people to the group. These very general descriptions are meant only to give an idea of roles and position.

In a network position can be more precisely defined. The position of a node is dependent on the incoming and outgoing links and the nodes it is linked to. In one definition[88], nodes v_i and v_j are said to have equivalent positions if they are linked to the same nodes (with the same direction links). Other definitions are also possible, based on link structure, isomorphic properties and other criteria. For example, in Figure 6.1, using automorphic equivalence, one could say that nodes v_1 and v_2 are equivalent because they are pointed from the same node and both have two nodes that they point to. They are however not equivalent to node v_3 because it only points to a single node.

Roles, too, can be calculated directly from the network structure. In recent publications, nodes can be assigned different roles depending on their connections and those of its neighbors. Many researchers distinguish between nodes that are assigned the role of core and those that are bridges[40]. Cores are those nodes which are part of a densely linked portion of the network, while bridges are those nodes that connect these dense portions of the network.

Figure 6.2 shows a canvas where nodes can be plotted based on their relative values of degree and rawComm. The canvas is divided into four quadrants and labelled as the roles the nodes in those quadrants would assume.

Nodes with a high degree but low rawComm, display a tendency to have many friends but all in the same community (or few communities) so are labelled *big fish*. *Ambassadors* have many friends in many different communities, *bridges* do not have so many friends by they are in different communities and *loners* are the nodes with few friends in a small number of communities [78].

6.2 Homophily and assimilation

The expression "birds of a feather, flock together", encapsulates the idea behind the two principles of *homophily* and *assimilation*. We are not surprised to see people with similar traits spending time with each other. This also happens in networks of nodes other than people. Blog posts can link to other blog posts and so be considered a network. It is not unusual for similar blogs to link to each other. The same thing happens of course in bibliographic data sets.

While seeing similar people together is not unusual, sometimes we would be interested in why they are linked. Homophily and assimilation are really explanations for the behavior. Homophily is the principle that people who have certain traits look for friends with the same traits. A person who enjoys participating in sports often times searches for friends that also like sports. Assimilation is where someone is already friends with someone and begins to adopt the traits of their friend [70]. You may have become friends with someone out of convenience (e.g., they lived next door), and as you get to know each other, you may start to become interested in their interests. Many people can relate to both of these principles.

The practical implications of studying homophily and assimilation are to offer advice to counselors, parents and health professionals. Generations of parents have urged their children to spend time with "good" kids. In some cases that is justified. In Pearson, et al., they were able to show that the behavior sometimes depends on the trait. The study tracked 50 middle school girls in Scotland and found that the girls adopted cannabis through assimilation, but they sought out friends who were smokers according to homophily. Both homophily and assimilation were observed with alcohol use.

A similar concept is assortative mixing [67]. Links in a network can be considered assortative or disassortive, where assortative essentially means the nodes are similar. Networks formed with assortative links show a high degree of similarity between the properties of nodes that are linked. Consider a politically aware social network where conservatives are mostly friends with other conservatives and liberals are friends with other liberals. On the other hand, some networks have disassortative links, like a dating network, where the linked nodes are mostly between people of different genders.

Knowing the nature of links (assortative vs. disassortative), an analyst will look at the network differently. Clustering coefficient will have different meanings. In an assortative network, a high clustering coefficient means that similar nodes tend to "stick together". High clustering coefficient in a disassortive network means just the opposite, that nodes with different characteristics tend to group. Assortative mixing can also affect reciprocity and path-based metrics (diameter and some centrality metrics).

6.3 Strength of weak ties

In an influential 1973 paper, Mark Granovetter laid out the concept of the "strength of weak ties" [35]. In it, he defined the strength of a link (tie) between two nodes based on the common neighbors (cm) metric. The higher cm between v_i and v_j, the stronger the tie between them. His paper argues that weak ties can be more important in some circumstances than strong ties.

The typical examples of circumstances that favor weak ties are finding a new job and meeting a new love interest. In both cases, a person's social networks include close friends and family who are acquainted with their other friends and family members but they don't know many people outside of the person's social sphere. Say that node v_i has a strong tie (high cm) with v_j. Then v_j would be unlikely to know many people that v_i does not already know. However, say that v_i has a friend v_m that is a weak tie (low cm). This could be someone they met once at a conference or on vacation. The chances are much better that v_m could put v_i into contact with a new job opportunity or a potential love interest, than v_j could.

This theory has been very influential in sociology, business, politics and other areas by encouraging people to improve their lives by

exploiting these weak ties. Another implication involves information acquisition. In a social network a person with few weak ties will get information from their close-knit friends, meaning the information will be from a single point of view. Having more weak ties increases the likelihood of getting more diverse sources of information.

Chapter 7

Models of Network Formation

The metrics and techniques for analyzing networks presented in this book operate on a static representation of the network at a given point in time. Networks are by nature evolving, links are added and removed as are nodes. Attributes are modified and groups are continually changing. While there has been some work done on dynamic networks, they require many snapshots of the network which complicates the analysis and requires more storage and memory.

So while we study static networks it is important to remember that they are just snapshots of dynamic systems. It is helpful to understand the network evolution that explains how the network came to be. There have been several theories put forth that provide models for network behavior. Models gain acceptance based on two criteria: their simplicity and, more importantly, how well they explain characteristics of physical world networks.

Like probability distributions, network models are simplified, idealized mathematical explanations for the observances in the real world. A sample of student IQ's from a school or region may likely take the shape of a Gaussian or normal distribution. Electronic components are often said to follow an exponential distribution. Having a model allows us to say something meaningful about the sample and to make predictions. In a simple case, knowing a student's IQ can help counselors gauge her expected success in school.

Models for networks play a similar role. Having a model for a network can influence what tools we choose to use for analysis. They can also tell us what to expect in the network. Like probability distributions, calculating a few key metrics can allow us to declare the underlying model that we believe is responsible for generating the network.

For all of the models we will examine both the characteristics of networks that obey the model and generating functions. In Figure 7.1, the top drawing (a), shows the process for calculating the statistics that determine the characteristics of the network. The bottom drawing (b), shows the process of entering parameters to allow the generating function to create a new network. Generating functions are mathematical formulas or descriptions for creating networks. They are probabilistic so that generated networks are unique but take on the characteristics of the model for which they were designed.

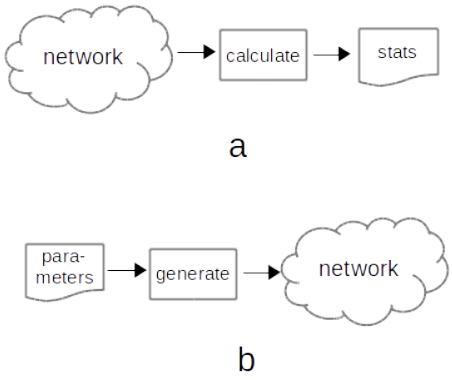

Figure 7.1: Characteristics and generation

7.1 Random

The random model for networks is the oldest of the models. They were first defined by Erdős and Rényi in 1959 [27]. A random network has the characteristics one would expect in a network where the links are placed randomly between nodes.

There are two common generating functions that are capable of creating networks with these characteristics. The first, *G(n,m)*, has the pa-

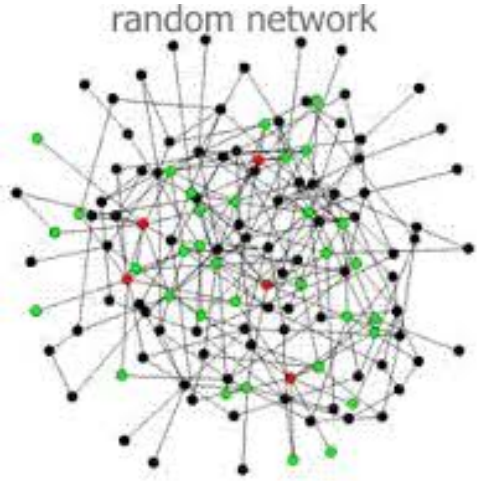

Figure 7.2: Random graph

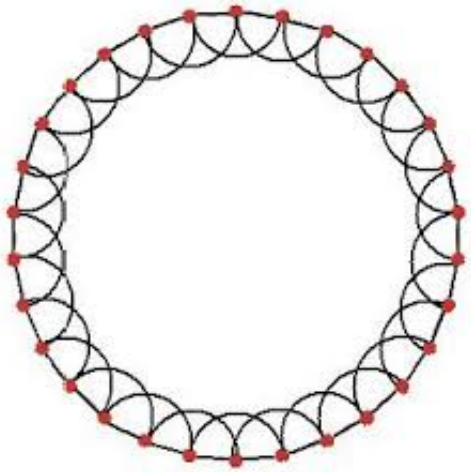

Figure 7.3: Regular circular network

rameters n (number of nodes) and m (number of links). The function operates by placing the m, links, one at a time, between two nodes chosen uniformly at random.

The second function, *G(n,p)*, is simple like the first function but it is sometimes preferred because it has some helpful mathematically provable characteristics. The function uses the parameters n and p (probability of link). For each possible link between all $n(n-1)/2$ pairs of nodes, a link is placed with the probability of p. For example, assume that $n = 10$ and $p = 0.1$. The function first considers the possible link between v_1 and v_2. A random number between 0 and 1 is drawn and if it is less than or equal to 0.1 a link is inserted between v_1 and v_2. The process continues for each of the node pairs.

Given a network that we believe to be random, specifically one generated by $G(n, p)$, there are a number of characteristics that this graph should have. The first is that the mean degree $c = (n-1)p$. This can be easily seen by dividing the total links (times 2 for each end of the link) by the number of nodes n, and we know that the number of links is $n(n-1)p/2$. The following properties can also be shown for random graphs [64]:

- the degree distribution follow a Poisson distribution.
- the average clustering coefficient $C = c/(n-1)$.
- the diameter is $a + \frac{\ln n}{\ln c}$ (where a is a constant).

There are two interesting things to notice about the properties listed. First, consider the clustering coefficient. With a medium sized network, say $n = 100,000$, and where the average degree $c = 100$, the clustering coefficient would be about $C = 0.0001$ which is quite small. Second, the diameter is also quite small. Take for instance, a random network the size of worlds population of 7 billion with an average degree of $1,000$. Disregarding the constant, this would mean there would be a diameter (maximum shortest path) of 3.33.... This is an advantage for random networks since short path lengths are normally desirable.

Random graphs can be popular with graph theoreticians for the mathematical properties described above. However, they are not as popular with network miners or analysts because only some are representative of real world networks and also because of the poor clustering.

7.2 Small world

In the late 1990's and early 2000's, a great interest in the study of large networks grew out of the availability of networks (such as the WWW), the increase in processing power of computers and the introduction of on-line social networks. Before this time, the random graph model was prominent. Now, however, researchers were looking for models that would better represent the characteristics that were evident in large real world networks.

In 1998, a sociology student and a mathematics professor, Watts and Strogatz, wrote a paper describing a new model, called small world, which better described the characteristics that they measured in many social networks [89]. The social networks in their observations, had

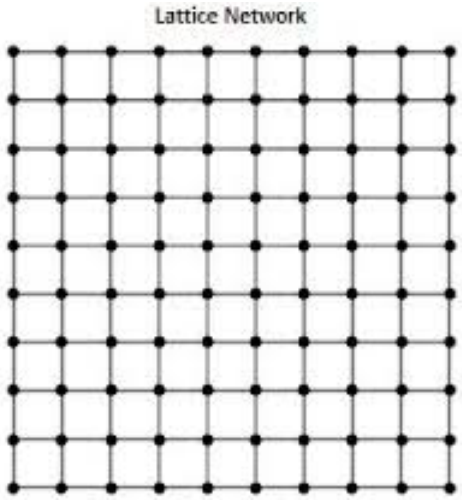

Figure 7.4: Regular grid network

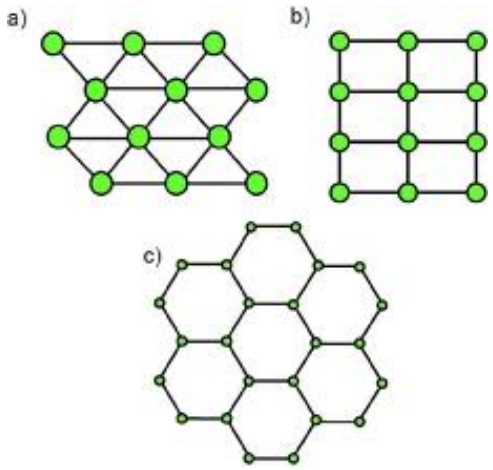

Figure 7.5: Regular misc. networks

the short average path lengths of random networks but in addition they also had very high average clustering coefficient.

Another class of graphs, regular graphs, where every node has the same degree and is connected to the k closest nodes. Because of the way the graphs are knit together, regular graphs have a very high clustering coefficient. For the same reason, though, they also have a very high average path length.

Watts and Strogatz were able to show that regular and random networks were at opposite ends of a continuum. On one end, the regular networks show zero randomness where the random networks at the other end are completely random. They suggested that one can generate networks starting with a regular network of n nodes, and using a probability parameter, p, rewire the links. When $p = 0$ no links are rewired and so you still have a regular network. When $p = 1$ all the links are rewired, resulting in a random network. For values of p between 0 and 1, the graph would be something between a regular and a random network.

Recall that regular networks have a high clustering coefficient and a high average path length, while random networks have a low clustering coefficient and a low average path length. Studies of real networks, exhibited high clustering coefficient and a low average path length. Watts and Strogatz found through experiments that the curves the two metrics appear similar to the ones in Figure 7.7. The dashed curve at the top is the average clustering coefficient and the lower dotted line is the average path length. They noticed that with values of p between 0.01 and 0.1, the generated network has the desired characteristics observed in real networks. They dubbed these networks small world networks because, like real social networks nodes are strongly connected to their neighbors and just a few hops away from any other node in the network.

7.3 Scale free

At about the same time that Watts and Strogatz were formulating small world networks, the researchers Barabási and Albert were also studying the characteristics of large networks [6]. What they noticed during their research is that many large networks had degree distributions that followed a power law distribution (see Figure 5.11). The degree distributions were fitted to the power law formula, $p_k = \frac{C}{k^\alpha}$ with

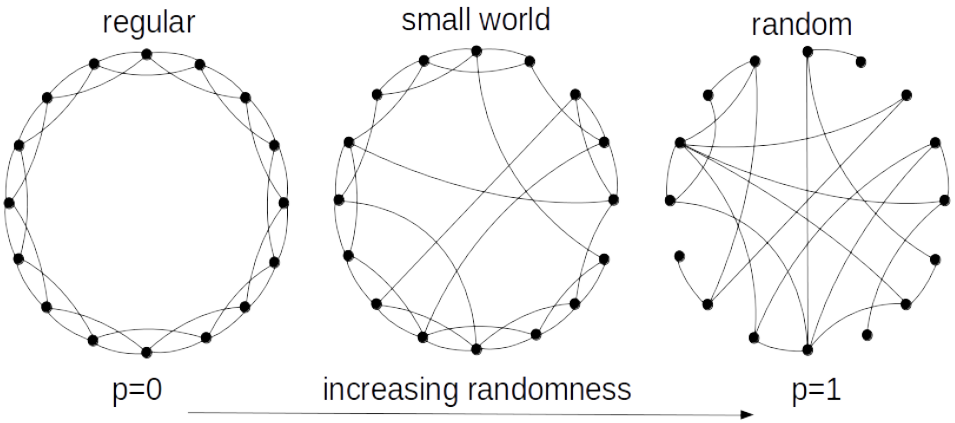

Figure 7.6: Regular - random continuum

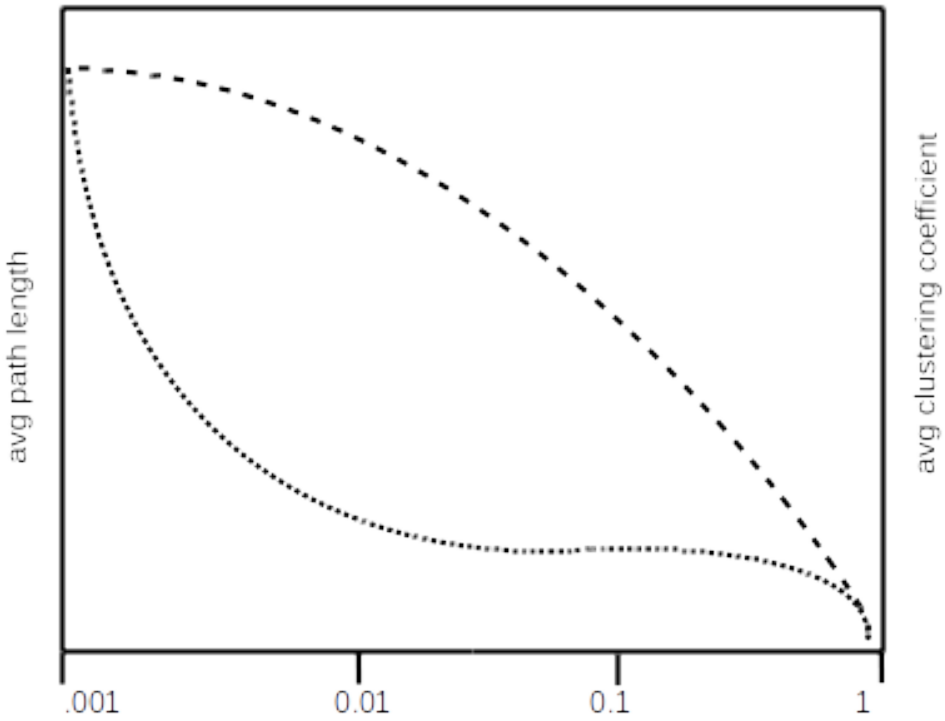

Figure 7.7: Small World Characteristics

$2 \leq \alpha \leq 3$. These networks are called scale-free.

As the plot and the formula for the power law show, the probability of a node have a small degree is much higher than having a high degree. Typically there are just a few nodes with very high degree and then it quickly falls with many more having a low degree. Imagine walking into a large room with two or three authoritative people surrounded by admirers. In such a situation, a new entrant to the room will often wish to make a connection with one of the "popular" people.

This analogy illustrates the preferential attachment approach to growing networks. Preferential attachment, sometimes referred to as "rich get richer" just means that the probability of a new node making a link to another node is proportional to the node's degree. There are many implementations of this generating function but a simple method is to create a m length vector of numbers filled in with node IDs, one for each link. For example, if node v_1 has just one link then 1 is copied into the vector just once but if v_2 has 12 links to other nodes, then 2 would be copied into the vector 12 times. A network that is created from a preferential attachment generating function will have a power law degree distribution.

There is no direct relationship between scale-free and small world networks. It is not unusual for a scale-free networks to also have the characteristics of small world nets. Conversely, it is often the case that small world networks also have the power law degree distribution. However, it is possible for a network to behave according to one of the models and not the other.

Chapter 8

Network mining techniques

For several decades, statisticians have applied their formulas to computer programming to automate them. This led to the areas of pattern recognition and machine learning. These areas combined mathematics, statistics and computer science to develop new algorithms used to make predictions and describe data sets. Starting in the late 1990's, with the growth of information, researchers interested in the application of these algorithms to large data sets forged the new area of data mining.

8.0.1 Data mining and network mining

Data mining is the search for hidden knowledge in large data sets. In general data mining applications do not provide exact answers to questions. Typically it is more like an expert who gives answers that are helpful and often correct but can occasionally be wrong. For example, imagine that we have a large data set containing the records of tax payers in the U.S. Say that we wanted to find the person who paid the most tax in the past year or the group of people who earned over $\$100,000$ and paid no tax. Both of these tasks are rather simple queries that a computer can execute exactly. In contrast, a typical data mining task would be, given the past information of taxpayers (both honest and cheaters), to predict tax cheats given new tax returns.

These type of tasks use labeled instances build models to predict new instances that are not labeled. In our analogy, taxpayers who are caught cheating are labeled as such and the others are labeled as not cheaters. The most well known data mining techniques are:

- classification – where the label of new instances are predicted from models built from past data.
- clustering – that groups the data according the similarity of the properties. For example, marketers are often interested in grouping people by demographics or other properties to better understand the underlying data.
- association analysis – that looks for frequently associated items within transactions. The results of association analysis is often seen in large on-line retailers who let you know that when people shop for the item you are looking for, they often also look at another suggested product.
- anomaly detection – which finds specific, unusual situations. Credit card fraud is an obvious example.

These techniques assume that the data is in the typical tabular format of rows and columns. While that is often the case with business, governmental and other stored data, data is also available in a number of other formats. For example, articles and stories (considered unstructured data), multimedia data like music and video, streams of data from sensors and news feeds, and network data from social networks and other applications.

With the growth of network data and the interest in data mining, new applications have been created to mine these large networks. The goal is the same, to find hidden knowledge, but in a different data format.

8.0.2 Sample network

To illustrate the concepts in this chapter, we will use a small sample network. To give a familiarity to the network we gave the nodes names based on the characters from the movie *Shrek*. However, all of the other attributes and the links have been invented by the authors. The intention is to provide a simple network with just enough data so that the techniques

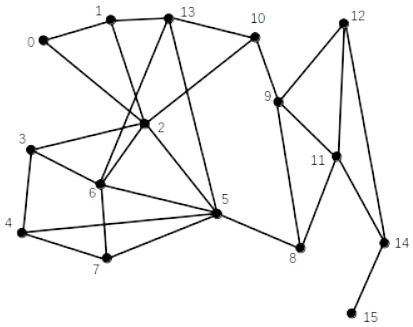

Figure 8.1: Weighted network

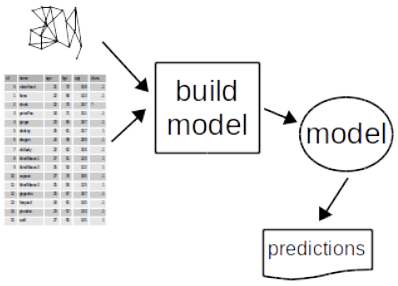

Figure 8.2: Node prediction process

id	name	age	hgt	wgt	cls
0	robinHood	21	70	168	-1
1	fiona	22	68	143	-1
2	shrek	27	61	187	?
3	peterPan	24	71	151	-1
4	ginger	23	66	187	-1
5	donkey	26	61	167	1
6	dragon	27	99	289	1
7	oldLady	22	62	166	-1
8	mouse1	27	51	120	1
9	mouse2	25	52	122	1
10	captain	23	70	186	-1
11	mouse3	25	50	125	1
12	geppetto	25	67	187	-1
13	farquad	24	61	145	-1
14	pinochio	23	57	133	-1
15	wolf	27	65	145	1

Table 8.1: Attribute data for sample network

can give us some meaningful results. Please don't be offended that your favorite character was omitted or that Donkey does not have a relationship with Peter Pan.

Figure 8.1 shows the network with labels for each node. The labels are the id which is listed in the attribute table. There are 16 nodes with a maximum degree of 6 and a minimum of 1. The average path length is 2.5 and the average clustering coefficient is 0.35. Fitting it to a power law curve gives us an exponent of 2. The reader With these statistics, the reader should decide if this network is a) small world and b) scale-free.

The attributes for this network are listed in Table 8.1. The characters range in age from 21 to 27, their heights are from 50 to 99 inches tall and their weights vary from 120 to 289. Notice that they also have labels (column labeled "class"). This is just an arbitrary designation and not to mean anything, but if it helps, notice that the $1's$ are characters based on non-human animals while the ones with a label of -1 are based on humans. Notice also that Shrek is labeled with a ?. We will predict his label in the section on node prediction.

8.1 Node prediction

Node prediction is also known as link-based classification. It is based on the classification problem of predicting the class of an object (which is simply an attribute of interest). A training set of objects whose class is known is given. A classifier is trained on the training set and used to predict the class of objects whose class is unknown. In a social network, for example, we may know the marital status of many of the members but not for all. The typical classification process uses input from a table of values and builds a model. The model can then be used to predict the class of new instances.

Traditional classifiers make a simplifying assumption that the objects are independent of each other. Researchers have recently begun to take advantage of the clearly defined relationships (links) within networks to improve classification. In the social network example above, the marital status of a person can potentially be inferred from the marital status of his/her friends. The challenge for node prediction (or link-based classification as it is often called) is integrating the attribute and link data. Figure 8.2 illustrates the process for node prediction. The main difference between it and traditional classification is that the input is based on the graph as well as the attribute data.

Using the attributes of neighbors has been

shown to actually be detrimental in some cases [14], however, using the class of the neighbor has been shown to be helpful [14, 55]. A related challenge is to recognize and utilize the structures inherent in the network. The study by Yang et al. [91] identified the existence of certain regularities in networks. For example, some networks exhibit encyclopedia regularity where nodes of one class link to nodes of the same class. We are more likely to be successful with node prediction using networks that have encyclopedia regularity.

Some researchers have concentrated on utilizing a local approach to node classification. For example, Chakrabarti et al. [14] have developed a technique for Web page categorization that exploits link information in a small neighborhood around the Web pages. They showed that, by using both the attributes of a node and the class of its neighbors, the error rate of an attribute-based classifier can be reduced up to 70%. In another work by Lu and Getoor [55], two classifiers were trained, one on the attribute data and the other using neighborhood class statistics of neighbors. They showed that the combined classifiers result in improved predictions.

Probabilistic models have also been used for link-based classification. Taskar et al. [84] proposed a probabilistic relational model using conditional Markov networks. Imagine the probabilities in the sample network. For example, lets say there is a 90% probability that a character has a class of -1 if they are younger than 24. Now assume that their class also has a similar probability with the age of each of their neighbors. The Taskar model builds a model that is itself a network to represent the probabilities of the original network. They showed that the collective classification of multiple related entities can be inferred from the learned model. Similarly, Neville and Jensen [63] proposed a generative model that simultaneously learns the hidden communities and the conditional probabilities associated with them. There have also been models proposed that are based on collective classification where the unlabeled instances are continuously relabeled during the model-building process.

Link-based classification can be extended to make use of the temporal information of an evolving network. For example, the class distribution of the nodes may change over time, and thus, can be exploited to improve the prediction. In some applications, a node can be assigned to multiple classes (e.g., a Web page having multiple tags or a gene with more than one functional classes). Therefore, another new direction is to learn all the classes associated with a given node, a problem that is known as multi-label learning. In another direction, while some of the above approaches have shown improvements by using the class information of neighbors, the information will be less helpful for some nodes than for others. Scripps et al. [77] have shown that the role that a node plays in the network can guide the classifier to use the neighborhood information when it is likely to help. For example, neighborhood information is less predictive for nodes linked to many communities.

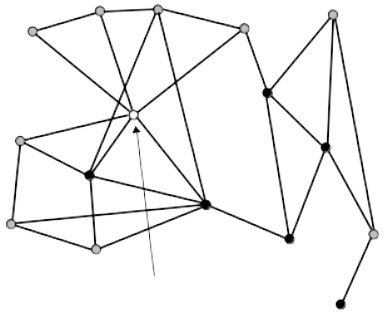

Figure 8.3: Node prediction example

There have been many applications of link-based classification. First, there is prediction time to churn [3], the time it takes for a customer to unsubscribe from a service or vendor. This is helpful for on-line retailers and information services to predict revenue streams and improve customer retention. In another [86], topics for a social media site are classified by sentiment or in other words how people feel about the topics (positive or negative). There are also examples of social network recommender systems [39] – using the network to recommend products or information for users and opinion mining systems [11].

To illustrate node prediction refer to Figure 8.3 where all of the nodes are labeled except node v_2 (Shrek) with an id of 2. The dark gray nodes are -1 while the white nodes are 1. Notice from Table 8.1 that nodes labeled -1 appear to be younger than the ones labeled 1. However there does not seem to be such a clear pattern to height and weight. A traditional classifier would probably classify Shrek as a 1 based on his age. However, when the Lu

& Getoor model is applied to the data, it predicts Shrek to be -1 based on his association with other characters labeled -1.

8.2 Link prediction

The *link prediction* problem can be stated as follows: Given a network, can we infer the node pairs that are likely to be linked together. Link prediction is applicable to static networks (to infer missing links in an incomplete network) or dynamic networks (to predict new interactions that will occur in the near future). Examples of link prediction problems include helping law enforcement to detect covert ties between criminal suspects or identifying future collaboration between researchers.

Link prediction is a challenging problem due to the sparsity of many networks. Predicting which non-linked node pairs will become linked has so far yielded very low accuracies [53]. Rattigan and Jensen [75] have shown that this is due to the skewed class distribution—as networks grow and evolve, the number of non-linked pairs increases quadratically while the number of linked pairs often grows only linearly. Recall from Chapter 6 that there is a tendency for individuals to establish friendship ties with others who have similar interests (attributes) [44]. In addition, individuals may also become friends because they share common friends (link structure) or belong to similar groups (communities). Integrating these different sources of information to improve link prediction is a challenge.

Liben-Nowell and Kleinberg [53] compared a large number of graph metrics as link predictors. They tested the metrics on bibliographic data sets using only the link structure and ignoring the node attributes. The results were not very predictive but one of the lessons from this work is that simple metrics like common neighbors and the Jaccard metric are just as predictive as the more complex ones. This work has been expanded to include both link and attribute data [38, 53] by using binary classifiers. Another approach is to use probabilistic generative models, where the goal is to learn the joint probability density of the nodes, links, subgroups, etc., and to predict the missing links by applying Bayes theorem [63, 85]. Rattigan and Jensen [75] proposed a variation to the problem known as anomalous link discovery, where the goal is to find links that are anomalous. Recent works have also considered the changes in the network over time. Potgieter et al. [73] combined the metrics from the Liben-Nowell study with temporal metrics such as return, moving average and recency. In another work by O'Madadhain et al. [68], a time evolving probabilistic classifier is constructed from training data sampled over many time periods. Hanneke and Xing [37] developed an extension of the Exponential Random Graph Model to account for the evolution of networks over time. A recent study by Backstrom et al. [4] on the evolution of communities in large social networks suggested that community structures and link formation are closely related. Making use of latent community structures for link prediction is another possible direction for future research.

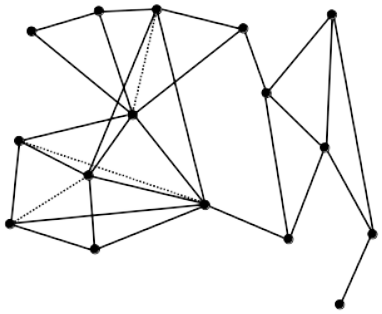

Figure 8.4: Link prediction example

For the sample, Shrek, network, using the simple, common neighbors metric, to predict links, one can see in Figure 8.4 that the most likely links are those between Shrek and Farquad, Peter Pan and Donkey and between the Gingerbread man and Dragon. You can confirm that all of those potential links have three common neighbors, which is the highest in the network.

8.3 Ranking

Ranking is the process of creating a total ordering of the nodes in a network. The rank of a node reflects the measurement of some particular structural property of the network, with respect to the node, which conveys a semantic meaning such as importance, popularity, authority, etc. As an end in itself, rankings can also be used to look for well-connected or *central* nodes in a network. Applications that use ranking include web search and community finding.

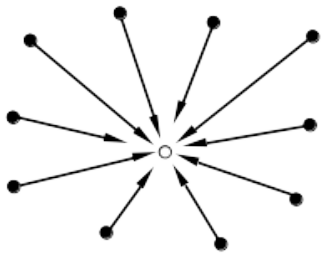

Figure 8.5: Unscrupulous web page (white) attempts to raise its ranking by creating dummy pages (black) to link to it.

Ranking can be done using centrality measures [87] like degree, closeness and betweenness. Another popular ranking method for large directed networks like the World Wide Web is the eigenvector method [48, 69] from Chapter 5. In this method a node's rank is the sum of the ranks of its incoming neighbors. Google's PageRank is an example of this ranking method. Nodes with a high rank are said to be authoritative as many other nodes refer to them. This method of ranking effectively stifles the problem of manipulation. In the World Wide Web, for which this was proposed, unscrupulous Web hosts would create fake Web pages linked to their main page to raise its rank (Figure 8.5). However, since these bogus web pages themselves have a low rank it does not increase the main page's rank very much. Unfortunately, this formulation has problems with graph cycles. The rank for nodes in a cycle will grow unabated. Page and Brin [69] solved this problem by adding a decay factor E yielding the equation $\alpha x = Ax + E$. Other approaches, similar to PageRank, include the algorithms HITS [48] and SALSA [50].

In addition to centrality measures, nodes can be ranked using other graph metrics based on their community belongingness. Guimera et al. [36] introduced a metric called participation coefficient, which measures to what degree a node participates in other communities. Their approach requires the communities in a network to be identified first using an algorithm that optimizes a modularity function of the network partition. As a consequence, the ranks of the nodes are sensitive to the choice of community finding algorithm. Also, rawComm from Chapter 5 can assign ranks and roles to nodes without applying a community finding algorithm.

8.4 Influence maximization

Closely related to ranking is the technique of *influence maximization* (also known as *diffusion of innovation*), which is important in the areas of epidemic spread and viral marketing. The goal is to find influential nodes – nodes that will spread their influence quickly through the network. In the context of social networks, influence can mean endorsing products or convincing others of a political idea. However, the concept can still be used in other kinds of networks. Influence is assumed to spread using a particular model of *diffusion*. In these models, nodes become activated (contracted a virus or bought a product) and can, in turn, activate their neighbors.

Diffusion models include the families of threshold and cascade models. In the threshold models [34] a node becomes activated when a certain percentage of its neighbors become activated. Newly activated nodes under the cascade models [33] have a one-time chance to activate neighbors with a given probability. Most of the models are probabilistic in nature. Without probability (e.g. if nodes are activated with certainty) every graph component with an activated node would end up with all nodes activated. Using appropriate probabilities ensures that activated nodes will only activate some of their neighbors and that the spread will stop before the entire network is activated.

The problem then is to choose (i.e. activate) nodes that will themselves, activate as many nodes as possible. For example, in viral marketing, a company may want to offer a small number of free or discounted products to influential people in the hopes that they will inspire others to purchase the product.

One might first consider activating only the highest degree nodes to obtain the optimal solution. However, one can quickly imagine that if the high degree nodes are all neighbors, the spread of influence will be less than if lower degree nodes – which are more spread out – were chosen. For example, in Figure 8.7, to maximize the number of nodes activated, the selected node is likely the best choice even though it is not the highest degree. Another challenge is that the link information may not be reliable – for example, in an online network, links between users are easy to add but do not always reflect genuine friendship. Furthermore, given

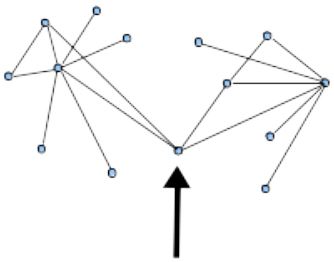

Figure 8.6: Choosing a node to maximize influence

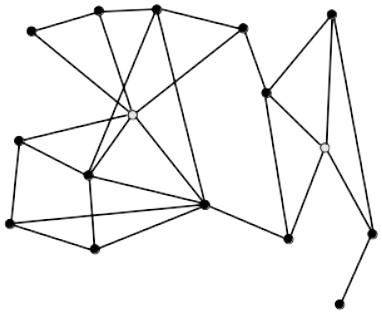

Figure 8.7: Using the greedy model of Kempe, et al., for influence maximization

the size of many real-world networks, simulating the activation process repeatedly to find the optimal solution is computationally expensive.

Kempe et al. [47] showed that the problem is NP-complete under the specific diffusion models of Independent Cascade and Linear Thresholds. They then propose a greedy strategy based on submodular functions [62], which guarantees a solution that is provably within 63% of optimal for these same models. In their experiments, the greedy strategy always performs better than the alternative strategies of selecting the nodes with the highest degree or lowest closeness scores. The greedy approach used by Kempe, et al., starts by finding the best node to activate using a brute force method. A node is activated and then the diffusion model is applied many times. After testing all of the nodes, the one that activated the most nodes is chosen. Then each of the remaining nodes is added to the first node and the simulations are run again to find the best node to add to the first one. This process continues until k nodes are chosen.

Various enhancements and improvements have been made to the greedy approach. Bharathi, et al. [10] extended the approach to account for multiple, competing innovations. The degree of a node is the number of outgoing links, i.e. the number of friends to which it is connected. While it is very fast to select the k nodes with the largest degree, this has been shown to be inferior to the greedy approach. However, Chen, et al. [16] used degree heuristics to improve the running time of the greedy algorithm. Narayanam, et al. [61], use the Shapley value from game theory as an heuristic to improve the running time of the greedy approach. The work in influence maximization is primarily concerned with maximizing only the raw number of nodes activated.

One can see from Figure 8.7 that Shrek and Blind Mouse 3 should be chosen to spread the influence to the maximum number of nodes in the sample network. This experiment was done using the Kempe greedy model and $10,000$ iterations.

8.5 Community finding

The technique of *community finding*, also called group detection [30], positional analysis [87, 24] or blockmodelling [87], is the process of placing nodes into groups in such a way that the nodes within a group are "similar" to each other and "dissimilar" to nodes in other groups. This is equivalent to clustering [82] and graph partitioning [45]. From the graph theory perspective, the problem of community finding is to remove links from the graph so that the remaining graph has the "desired" components.

Community finding is ill-posed; there is no agreed-upon metric for evaluation. One popular graph-based metric from Newman and Girvan [65], called modularity, is based on the fraction of links within a community to those between communities. An additional challenge to community finding is scalability. Networks such as the World Wide Web or online social networks can have millions, even billions, of nodes. Algorithms for community finding can be complex making them unusable for large networks. However, it is still a very helpful tool and research continues to find better algorithms and metrics.

The most common approach to community finding is to segment the entire network into disjoint groups, where each node is assigned to exactly one community. Traditional clus-

tering algorithms such as k-means, DBScan, Chameleon, etc. [82] can be applied to generate such communities. Graph partitioning algorithms are also applicable. For example, spectral clustering [80] divides a network into balanced components based on the eigenvalues of its Laplacian matrix. This approach is equivalent to finding a partitioning that minimizes the normalized cut criterion [22]. Karypis and Kumar [46] developed a multi-level graph partitioning approach that can accommodate different heuristic functions for coarsening, partitioning and refining the clusters. While these algorithms were not specifically designed for networks, their application is straightforward. An approach that was specifically designed for networks, from Girvan and Newman [32], uses the edge betweenness metric to remove edges iteratively. It is intuitively appealing since high betweenness edges would appear to be bottlenecks between communities; however, it is slow [74, 43]. Clauset, et al., proposed an agglomerative method of merging communities that optimizes modularity. This is a fast algorithm and has been improved to be even faster [**?**].

A variant of finding disjoint communities is to discover a hierarchy of communities. This approach allows the communities to be nested and organized as a tree structure called a dendrogram. Agglomerative hierarchical clustering methods such as single-link and complete-link can be used to find hierarchies in networks. More recently, Clauset et al. [18] proposed a method of extracting hierarchies based on maximum likelihood methods and Markov chain Monte Carlo sampling.

Algorithms that find disjoint communities are popular but do not allow for situations when nodes can belong to more than one community. This is often the case in social networks where, for example, an individual can join two or more communities. An extreme case is to use fuzzy clustering, where every node belongs to every community with an associated weight. Another overlapping clustering method, developed by Banerjee et al. [5], is based on a mixture model of exponential family distributions. One of the challenges of overlapping communities is excess overlap. Finally Scripps and Trefftz [**?**] proposed an algorithm, CHI, that uses a Kmeans-like approach to clustering nodes. CHI produces disjoint as well as overlapping communities, optimizing an objective function that allows the user to determine the type of communities of interest.

Some community finding methods do not try to completely cluster the entire network. Instead they form communities from a given root set of nodes. This approach is helpful when only a portion of the network is known or the network is large and a complete grouping is unnecessary. As examples, consider Web page search where a set of related pages can be useful or in a large bibliographic database, where one only wants to know the community of researchers to which an author belongs. A Markov chain approach by Gibson, et al. [31], specific to the World Wide Web, starts with a core set of pages, adds a fixed number of nearby pages, then forms the communities from the authoritative pages in the expanded set. Min-Cut has also been adapted [28] to find a community by using the targeted node as the source and adding a virtual sink connected to all nodes in the graph.

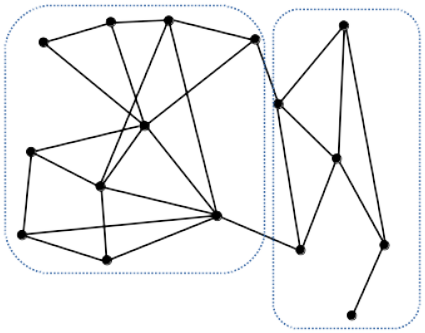

Figure 8.8: Two communities detected in the Shrek network

In the Shrek network, Figure 8.8 shows the two communities that were found using the CHI algorithm. The three blind mice, Pinochio, Geppetto and the wolf are in the community on the right and the rest are in the community to the left. Notice that only two links are "broken" – that is, they connect nodes in different communities. Another way to partition the nodes into two communities is to put the wolf into a community by itself and all the others in the other community. While this would result in only one broken link it is not a very interesting set of communities. While the different algorithms have different behaviors, most keep densely linked groups of nodes together in the same community.

Chapter 9

Security, Privacy and Trust

9.1 Security

9.2 Privacy

9.3 Trust

Bibliography

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[3] A. Backiel, B. Baesens, and G. Claeskens. Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, 2016.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[5] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. Mooney. Model based overlapping clustering. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[6] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288:50–59, May 200 3.

[7] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[8] P. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91, 2004.

[9] Graham Bennett and Kalemani Jo Mulongoy. Review of experience with ecological networks, corridors and buffer zones. In *Secretariat of the Convention on Biological Diversity, Montreal, Technical Series*, volume 23, page 100, 2006.

[10] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *Proceedings of WINE*, 2007.

[11] P. Bo and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 2008.

[12] Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):22, 2011.

[13] Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. Batch kernel som and related laplacian methods for social network analysis. *Neurocomputing*, 71(7):1257–1273, 2008.

[14] S. Chakrabarti, B. Dom, and P.Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the SIGMOD International Conference on Management of Data*, 1998.

[15] H. Chang and D.-Y. Yeung. Robust path-based spectral clustering. *Pattern Recogn.*, 2008.

[16] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[17] Robert M Christley, GL Pinchbeck, RG Bowers, D Clancy, NP French, R Bennett, and J Turner. Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology*, 162(10):1024–1031, 2005.

[18] A. Clauset, C. Moore, and M. E. J.Newman. Structural inference of hierarchies in networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML), Workshop on Social Network Analysis*, 2006.

[19] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

[20] A. Davis, B. B. Gardner, and M. R. Gardner. *Deep South*. The University of Chicago Press, 1941.

[21] James Decraene and Thomas Hinze. A multidisciplinary survey of computational techniques for the modelling, simulation and analysis of biochemical networks. *j-jucs*, 16(9):1152–1175, may 2010.

[22] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[23] Twitter marketing stats: 2015 - infographic. http://www.digitalinformationworld.com/2015/01/twitter-marketing-stats-and-facts-you-should-know.html.

[24] P. Doreian, V. Batagelj, and A. Ferligoj. Positional analysis of sociometric data. In P. Carrington, J. Scott, and S. Wasserman, editors, *Models and Methods in Social Network Analysis*. Cambridge, New York, 2005.

[25] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[26] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook friends: Social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[27] P. Erdös and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[28] G. Flake, K. Tsioutsiouliklis, and R. Tarjan. Graph clustering techniques based on minimum cut trees. Technical report, NEC, Princeton, NJ, 2002.

[29] J. Gao, P.N. Tan, and H. Cheng. Semi-supervised clustering with partial background information. In *Proceedings of SDM'06: SIAM Int'l Conference on Data Mining*, 2006.

[30] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7:3–12, 2005.

[31] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 1998.

[32] M. Girvan and M. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci*, 99:7821–7826, 2002.

[33] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing*, 01, 2001.

[34] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83, 1978.

[35] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[36] R. Guimerà, M. Sales-Pardo, and L. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, 3:63–69, 2007.

[37] S. Hanneke and E. Xing. Discrete temporal models of social networks. In *Proceedings of the 23rd International Conference on Machine Learning Workshop on Statistical Network Analysis*, 2006.

[38] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Proceedings of SDM'06: SIAM Data Mining Conference Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[39] J. He and W. Chu. *A social network-based recommender system (SNRS)*. Springer, 2010.

[40] Shaobin Huang, Tianyang Lv, Xizhe Zhang, Yange Yang, Weimin Zheng, and Chao Wen. Identifying node role in social network based on multiple indicators. *PloS one*, 9(8):e103733, 2014.

[41] Instagram statistics. http://opticalcortex.com/instagram-statistics/.

[42] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *IEEE computer*, 29(3):31–44, 1996.

[43] J.R.Tyler, D.M. Wilkinson, and B.A.Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Proceedings of the First International Conference on Communities and Technologies*, 2003.

[44] D. Kandel. Homophily, selection, and socialization in adolescent friendships. *The American Journal of Sociology*, 84:427–436, 1978.

[45] G. Karypis and V. Kumar. Analysis of multilevel graph partitioning. In *ACM/IEEE conference on Supercomputing*, 1995.

[46] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20:359–392, 1999.

[47] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[48] J. Kleinberg. Sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[49] D. E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, 1993.

[50] R. Lempel and S. Moran. Salsa: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19:131–160, 2001.

[51] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2005.

[52] Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

[53] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, 2003.

[54] livejournal download. http://download-lj.livejournal.com/.

[55] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of the 20th International Conference on Machine learning*, 2003.

[56] Richard Martin. Are cell phones replacing landlines. *Information Week*, 2007.

[57] Seth A Marvel, Travis Martin, Charles R Doering, David Lusseau, and Mark EJ Newman. The small-world effect is a modern phenomenon. *arXiv preprint arXiv:1310.2636*, 2013.

[58] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[59] Frank Müller and Frank Thiesing. Social networking apis for companies - an example of using the facebook api for companies. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 120–123. IEEE, 2011.

[60] Jussi Myllymaki. Effective web data extraction with standard xml technologies. *Computer Networks*, 39(5):635–644, 2002.

[61] R. Narayanam and Y. Narahari. A shapley value-based approach to discover influential nodes in social networks. *Automation Science and Engineering, IEEE Transactions on*, 8(1):130 –147, jan. 2011.

[62] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[63] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005.

[64] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[65] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, Feb 2004.

[66] M. E. J. Newman. Word adjacencies. *Phys. Rev. E 74*, 2006.

[67] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[68] J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations*, 7:23–30, Dec 2005.

[69] L. Page, S. Brin, R. Motwani, and T. Winograd. Pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[70] M. Pearson, C. Steglich, and T. Snijders. Homophily and assimilation among sport-active adolescent substance users. *Connections*, 27:47–63, 2006.

[71] 6 new facts about facebook. http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

[72] Sanjukta Pookulangara and Kristian Koesler. Cultural influence on consumers usage of social networks and its impact on online purchase intentions. *Journal of Retailing and Consumer Services*, 18(4):348–354, 2011.

[73] A. Potgieter, K. April, R. Cooke, and I. O. Osunmakinde. Temporality in link prediction: Understanding social complexity. *Journal of Transactions on Engineering Management*, 7, 2006.

[74] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. In *Proceedings of the National Academy of Sciences*, 2004.

[75] M. Rattigan and D. Jensen. The case for anomalous link discovery. *SIGKDD Explorations*, 7, 2005.

[76] J. Scripps and P. N. Tan. Clustering in the presence of bridge-nodes. In *Proceedings of SDM'06: SIAM Int'l Conference on Data Mining*, 2006.

[77] J. Scripps, P. N. Tan, and A-H Esfahanian. Exploration of link structure and community-based node roles in network. Technical report, Michigan State University, 2007.

[78] J. Scripps, P. N. Tan, and A-H Esfahanian. Node roles and community structure in networks. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Joint Workshop on Web Mining and Social Network Analysis*, 2007.

[79] J. Scripps and C. Trefftz. Discovering influential nodes in social networks through community finding. In *9th International Conference on Web Information Systems and Technologies*, 2013.

[80] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 22(8), August 2000.

[81] Tanmay Sinha. Supporting mooc instruction with social network analysis. *arXiv preprint arXiv:1401.5175*, 2014.

[82] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, Inc., 2005.

[83] C. Tantipathananandh, T. Y. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

[84] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02)*, 2002.

[85] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems Conference (NIPS03)*, 2003.

[86] X. Wang, F. Wei, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *20th ACM international conference on Information and knowledge management*, 2011.

[87] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[88] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[89] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, pages 440–442, Jun 1998.

[90] Wikipedia downloads. https://dumps.wikimedia.org/.

[91] Y. Yang, S. Slattery, and R. Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18, March 2002.

[92] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.