0 Introduction

net·work: (*n.*) an interconnected or interrelated chain, group, or system.

Imagine a world where people have no friends. Where roads never intersect. Where computers are not interconnected. This world without networks would be a very sad and boring place, where nothing happens — and even if something happened, nobody would know. Such a world is unimaginable, because our life is completely defined by networks: relationships, interactions, communications, and the Web. Biological networks governing the interactions between genes in our cells determine our development, neural networks in our brain make us think, information networks guide our knowledge and culture, transportation networks allow us to move, and social networks sustain our life.

Networks are a general yet powerful way to represent and study simple and complex interactions. This book explores the study of networks and how they help us understand the patterns of connections and relationships that shape our lives. In essence, a network is the simplest description of a set of interconnected entities, which we call *nodes*, and their connections, which we call *links*. The network representation is so general and powerful because it strips out many details of a particular system and focuses on the interactions among its elements. Networks are thus used to study widely diverse systems. Nodes can represent all sorts of entities: people, cities, computers, Web sites, concepts, cells, genes, species, and so on. Links represent relationships or interactions between these entities: friendships among people, flights between airports, packets exchanged among computers on the Internet, links between Web pages, synapses between neurons, and so on.

Before we introduce the basic concepts, definitions, and nomenclature about networks, let us explore a few examples of social, infrastructure, information, and biological networks. Data for all the examples presented here is available on the book's Github repository.[1] The networks on which we focus in this book tend to be large, even though one can learn a lot from studying smaller systems, such as social networks built from surveys or interviews. In these cases it is meaningful to examine individual nodes and connections in great detail, whereas analyses of

[1] `github.com/CambridgeUniversityPress/FirstCourseNetworkScience`

large networks tend to focus on macroscopic properties, classes of nodes and links, typical behaviors, and anomalies.

0.1 Social Networks

A social network is a group of people connected by some type of relationship. Friendship, collaboration, romance, or mere acquaintance are all examples of social relationships that connect pairs of people. When we talk about a social network, we typically think of a particular type of relationship. A person is represented by a node in the social network, and the relationship is represented by a link between two people. The network is therefore a representation of the relationship. It allows us to talk about the relationship, to describe it and analyze it at a level that goes beyond a pair of people.

There are many different types of social networks, and they are important to study. Health workers analyze networks of sexual relationships to find ways to combat the spread of sexually transmitted diseases. Economists study job referral networks to address inequality and segregation in labor markets. And scientists inspect coauthorship networks in scholarly publications to identify influential thinkers and ideas.

These days we use online social networking sites to keep track of our social ties. Platforms like Facebook and Twitter allow us to keep in touch with many people — partners, friends, colleagues, and acquaintances, sometimes in the hundreds — and communicate with them conveniently, irrespective of distance. Figure 0.1 illustrates a familiar network, a portion of the Facebook social graph. In this network, nodes are people with a Facebook account at a US university, and connections may represent different types of relationship, from real friendship to mere acquaintance. Just looking at the network visualization reveals something about the underlying social structure. Some people have more connections; we represent this by making the corresponding nodes larger and darker. These might be popular students, teachers, or administrators. We also notice that the network is roughly divided into two parts. The data is anonymized so we cannot tell for sure, but a possible interpretation is that the larger sub-network comprises mostly undergraduate students, and the smaller one includes mostly graduate students. There are connections between nodes in the two groups, but not as many as among nodes within each group. In other words, undergraduate students are more likely to be friends with other undergrads than with grad students. Later we will introduce formal names for all these observations, which are typical of most social networks.

The availability of data from online social networks is very exciting for scientists. We can study human interactions at a scale and resolution that was never possible in the past: who befriends whom, who pays attention to what, who likes what, what gets recommended, and how this information propagates through the network. This data provides us with an unprecedented opportunity to discover, track, mine, and

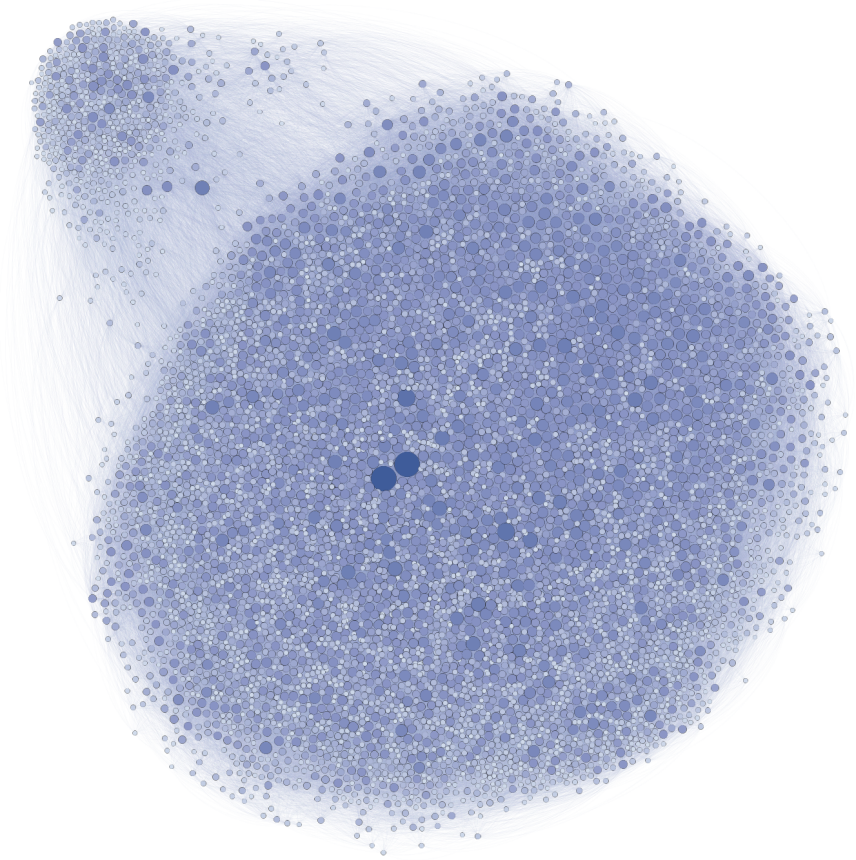

Fig. 0.1 Visualization of a network of Facebook users at Northwestern University. Nodes represent people, and links stand for Facebook friend connections.

model what people do. Like the telescope gave us a first view of distant planets and stars, and the microscope allowed to peek into living tissue and micro-organisms, social media are enabling the study of social systems and human activity. However, as exciting as these opportunities are to researchers, they don't come without risks of abuse. Online interactions expose our private personal information. We've all heard stories about employers finding embarrassing pictures of prospective employees, or scandals related to hackers and political organizations amassing data about millions of users. The dangers can be subtle. Knowing a little bit of information about a large number of people can reveal a lot more than intended. Using data from Facebook, two MIT students found that just by looking at the gender and sexuality of a person's online friends, they could predict whether the person was gay. Online social networks also make impersonation easy to do and hard to detect.

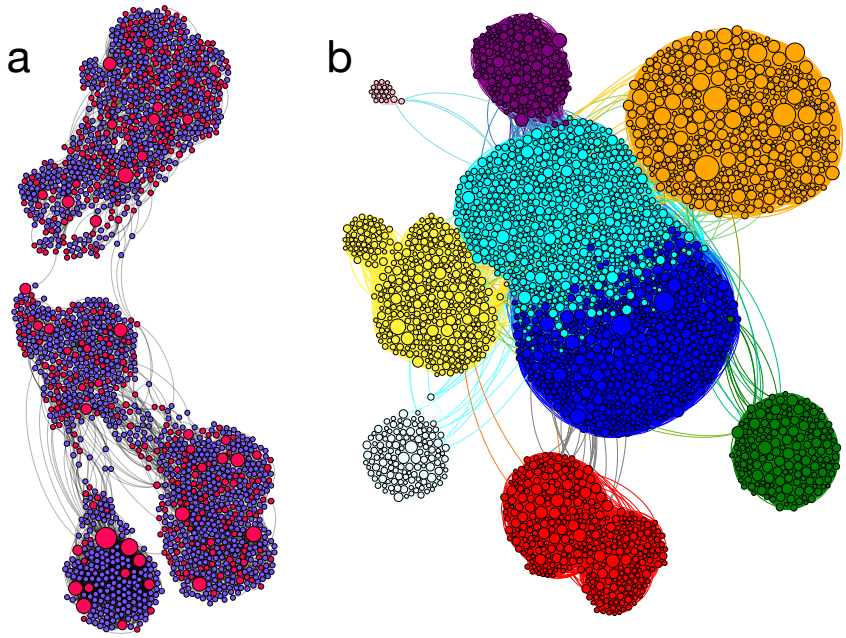

Fig. 0.2 (a) A movie-star network, based on a small sample of movies, actors and actresses from the Internet Movie Database. Nodes represent movies (blue) or actors/actresses (red). A link connects an actor or actress to a movie in which they starred. (b) A movie co-star network, based on a small sample of actors and actresses from the Internet Movie Database. A link connects two people who have co-starred in at least one movie. Colors represent film genres or languages/countries.

Social phishing is the technique of impersonating a victim's friend (as inferred from an online social network) to induce the victim to disclose sensitive information. Two Indiana University students demonstrated that they were able to obtain the secret passwords of 72% of victims in this way.

Data about a social network can be extracted from many sources. If we want to map human mobility patterns to improve urban transportation networks, we can collect call data from cell phones. If we want to map coauthorship among scientists, we can extract the names from a database of scientific publications; two coauthors of the same paper will be linked to each other. (This is not a trivial exercise, because several scientists have common names.) If we want to map the collaboration among movie stars, we can extract movie credits data from the Internet Movie Database (`IMDB.com`). Figure 0.2 illustrates two such networks. In one case, there are actually two kinds of nodes: movies and actors/actresses. We draw a link between an actress and a movie in which she has starred. In the other case, we focus on links between actors/actresses who have co-starred in movies. Although the de-

picted networks capture only tiny portions of the movie database, we again notice some clear patterns. Larger nodes have more connections, representing stars who acted in many movies. We also see that the networks are structured into several dense groups associated with periods, languages, or film genres: Hollywood (blue), Western (cyan), Mexican (purple), Chinese (yellow), Filipino (orange), Turkish and Eastern European (green), Indian (red), Greek (white), and adult (pink) stars in Figure 0.2(b). In Chapter 6 you will learn how to discover these groups and find out what they are about.

0.2 Communication Networks

In the Facebook and movie networks, links are reciprocal: you cannot friend someone on Facebook unless they agree, and you cannot star in a movie without being listed in the credits. Not all social networks have reciprocal links, however. For example, Twitter is a popular social network with links that are not necessarily reciprocal: Alice can follow Bob without Bob necessarily following Alice back. As a result, the relationships captured by the Twitter network is not friendship; you follow someone to see what they post. When you retweet a post, your followers see it. This is a good way to share information broadly, so Twitter is a social network mainly aimed at spreading information — a communication network. The retweet network in Figure 0.3 illustrates the spread of political messages during a US election. Larger nodes are those with more outgoing links, because how many times users are retweeted by others is a way to measure their influence. You probably noticed immediately a more striking pattern: conservative users (red nodes) mostly retweet messages from other conservatives, while progressive users (blue nodes) similarly share progressive content. In fact, such preferential patterns of the social connections allow us to guess a person's political leaning with high accuracy. This property, called *homophily*, will be discussed in Chapter 2; the algorithm for inferring political preference from the network's structure will be presented in Chapter 6.

Networks like Twitter let us trace the diffusion of hashtags and news, observing how ideas and cultural concepts spread from person to person. But social media are also used to spread misinformation, which is unknowingly passed on by gullible users. Using fake news sites and automated or semi-automated accounts known as "social bots," a malicious entity can cheaply and effectively generate and amplify a disinformation campaign, either for political purposes or to monetize traffic through ads. In recent years we have observed a sharp increase in these types of manipulation of social media on a global scale. If one can control what information people see online, one can manipulate their opinions. This is a threat to democracy in many countries, because without well-informed voters one cannot have free elections. Academic researchers and industry engineers are working hard to develop

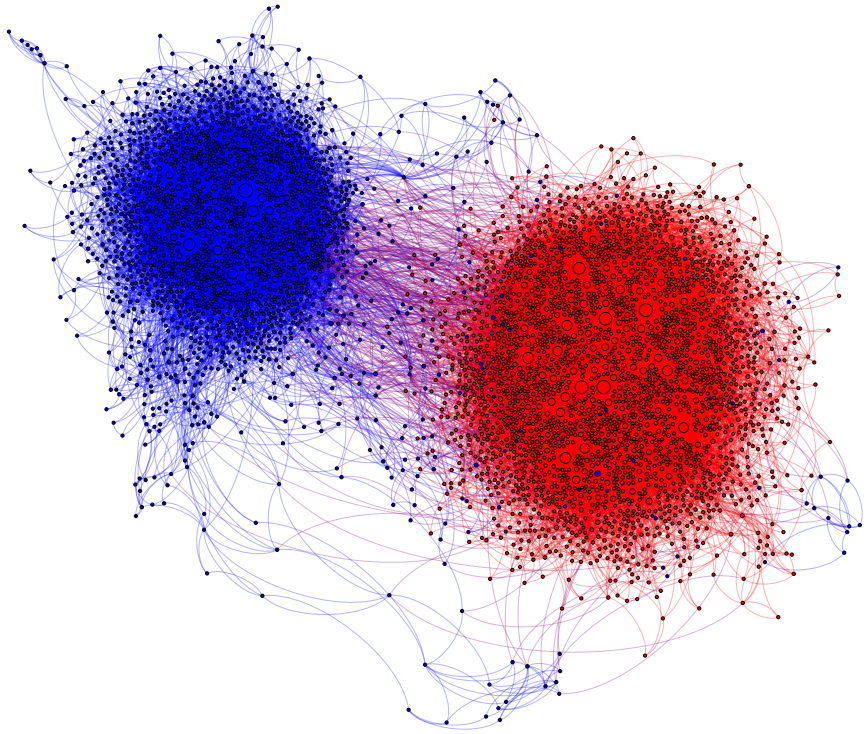

Fig. 0.3 A retweet network on Twitter, among people sharing posts about US politics. Links represent retweets of posts that used hashtags such as `#tcot` and `#p2`, associated with conservative (red) and progressive (blue) messages respectively, around the 2010 US midterm election. When Bob retweets Alice, we draw a directed link from Alice to Bob to indicate that a message has propagated from her to him. The direction of the links is not shown.

countermeasures. Understanding the structure and dynamics of the networks that enable the spread of information is a critical component of these efforts.

The social links in Twitter are in place before a user generates a post, which is typically broadcast to all of the user's followers. In email, just like in social networks, nodes are people. However, each message is intended for one or more specific recipients. Links are based on the messages exchanged. Email does not depend on a particular platform; the protocol is open and distributed, so that no single organization controls all of the traffic. As a result, email is still among the most widely used communication networks. Figure 0.4 illustrates an example of an email network. Again, links are directed from the sender to the receiver of an email, indicated by arrows. Node size and color represent two different features: number of incoming and outgoing links, respectively: a larger node receives emails from more

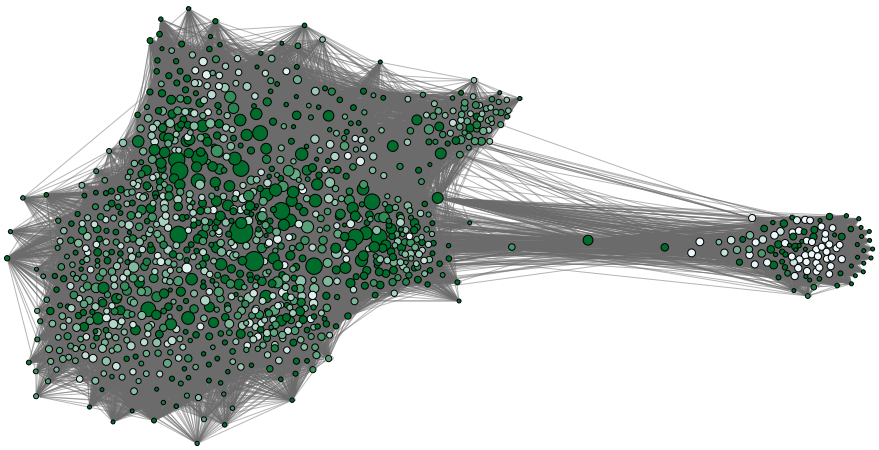

Fig. 0.4 A network based on a database of emails generated by employees of the Enron energy company. The data was acquired by the US Federal Energy Regulatory Commission during its investigation after the company's collapse in 2001. At the conclusion of the investigation, the emails were deemed to be in the public domain and made publicly available for historical research and academic purposes. Only a small portion of the central core of the network is shown. The direction of the links is shown by arrows.

people, and a darker node sends emails to more people. The fact that larger nodes tend to be darker and vice versa tells us that there is a correlation between sending and receiving emails.

0.3 The Web and Wikipedia

The Web is the largest information network. While it is now used to provide all kinds of services, it was originally just a network of documents (pages) connected by "hyperlinks," or clickable links. In the early 1990s, Tim Berners-Lee wanted to simplify access by scientists to information about high-energy physics experiments at the European Organization for Nuclear Research (CERN) near Geneva. He came up with three key ideas: (1) a naming system for pages, the Uniform Resource Locator (URL); (2) a simple language for writing documents, called HyperText Markup Language (HTML), including hyperlinks from one page to another; and (3) a simple protocol called HyperText Transfer Protocol (HTTP) for clients (browsers) to talk to servers. With these three components, the Web was born. Berners-Lee even implemented the first Web server and browser software to download pages and media from servers by clicking on links. We can actually see two networks at play here: the static "link graph" made of a snapshot of Web pages and links at a given time, and the dynamic traffic network emerging from people navigating the Web.

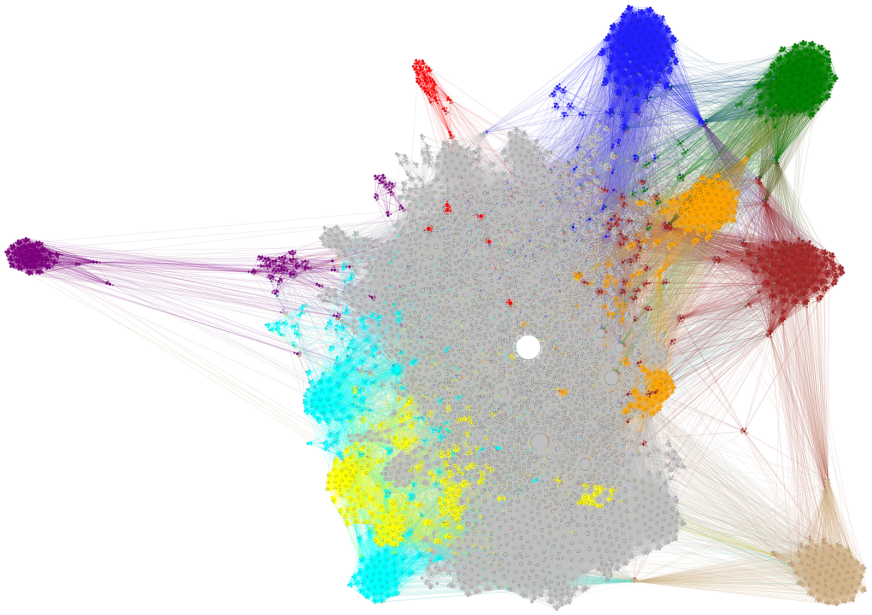

Fig. 0.5 A portion of the Wikipedia information network. Nodes are articles about math. We only consider links among Wikipedia articles, and disregard links to external pages. Node size is proportional to article importance, and colors highlight communities discussed in the text.

To paraphrase the classic philosophical riddle, if there is a link between two pages but nobody clicks on it, is it really part of the Web? The answer of course depends on which of the two networks we are thinking about when we say "Web." In later chapters we will spend more time exploring both of these information networks.

The Web is too large to visualize even a small portion of it in a meaningful way. Let us focus on Wikipedia, which is a network of pages (articles) on a single website. The Wikipedia is a collaborative encyclopedia edited by thousands of volunteers around the world, and it is one of the most popular destinations on the Web. There are versions of Wikipedia in many languages, so let us focus on the English one. Still, the English Wikipedia is a huge network with millions of articles (and growing!). So let us focus on just a small subset of articles about math, shown in Figure 0.5. Here, node size represents *PageRank*, a measure of centrality that captures how important is an article based on other articles that link to it — something we will discuss in Chapter 4. For example, the large white node in the middle is the general article about *Mathematics*. Another feature of this network is the presence of a large "core" (gray) and several smaller groups. These groups are tightly connected clusters of articles on specific topics or branches of math. For example, articles about historical Greek (blue), Arab (green), and Indian (brown)

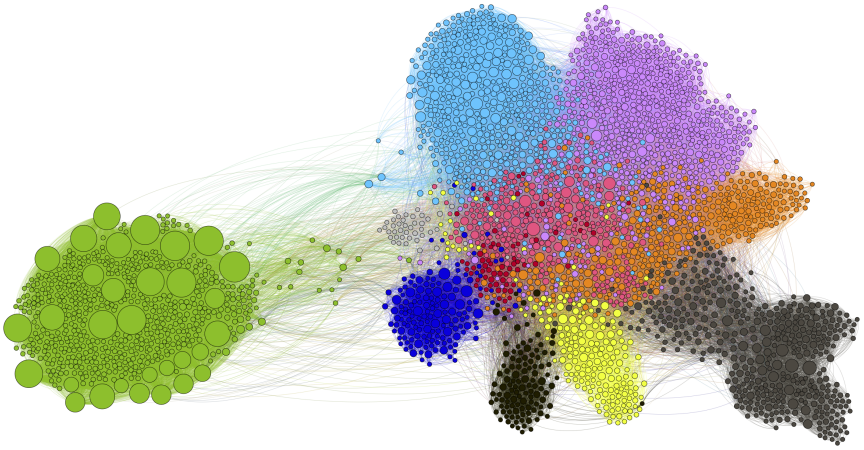

Fig. 0.6 A portion of the Internet router network. The map is a snapshot generated by the Center for Applied Internet Data Analysis (CAIDA.org) using tools that send out small packets of data (probes) between Internet hosts. Colors are assigned according to a community detection algorithm that identifies dense clusters reflecting the geographic distribution of routers. In Chapter 6 you will learn how to use this methodology to study what those clusters represent.

mathematicians; about contemporary Indian mathematicians (tan); about math and art (orange), statistics (cyan), game theory (yellow), mathematical software (purple), and pedagogical theory (red). We also observe several "bridge" nodes that connect multiple clusters. These features are found in many real-world networks.

0.4 The Internet

We often think of the Internet as a network of computers and other connected devices, but in reality it is a *network of networks*. In fact the word originates from *internetworking,* or connecting different computer networks through special nodes called *routers.* We can therefore observe the Internet at many levels: at the lowest level we have hardware devices that connect individual computers in the same local or wide-area network. These networks are connected by routers, so we can zoom out and think of the network of routers. If we zoom out further we find groups of networks managed by an Internet Service Provider (ISP). This organization decides its internal network topology (how routers are connected) autonomously, and therefore is also called "autonomous system" (AS). Special "border" routers connect one AS to another, forming what we call the AS network.

Figure 0.6 shows a small portion of the Internet router network. Although the

Internet has evolved without central control or coordination, ISPs follow local rules on how to connect their routers. They try to provide the best service at the lowest cost. Certain regularities emerge as a result. For instance, the portion of the Internet that carries the most traffic is often referred to as "backbone." The large telecommunication companies that manage the Internet backbone have a significant interest in preventing disruption, so they engineer their networks with a lot of redundancy. We thus observe a dense "core," with large routers connected to each other. As we move toward the "periphery" of the Internet — our home routers — the network is more sparsely connected. Such a hierarchical *core-periphery structure* is common in many different types of networks, and will be discussed in Chapter 2. In the router network depicted in Figure 0.6, the green cluster on the left appears well separated from the rest of the network. This is likely due to a bias in the probe methodology used to map these networks: most measurements were taken from the United States, and the routers in this cluster are located there. A related peculiarity is the presence of very large nodes in the green cluster, indicating routers with many connections. This may actually be a measurement error resulting from the same bias. In fact, a router can only have a limited number of connections due to hardware constraints. Let it serve as a reminder that if we use a flawed method to collect data about a network, its analysis may lead to wrong conclusions.

0.5 Transportation Networks

Another important class of networks concern various types of transportation. Nodes are locations: cities, road intersections, airports, ports, train or subway stations. These networks are very different from one another, however. Road networks, for example, evolve in a local fashion to minimize the distance traveled between nearby cities. This leads to the emergence of grid-like structures, in which most nodes have a comparable number of connections — say, four-way intersections. Figure 0.7 shows an air transportation network, which does not have a grid structure. The reason is that airlines try to minimize the number of hops between source and destination without adding costly direct flights between low-traffic airports. The simple solution is to add flights connecting airports to existing hubs. As a result, air flight networks display "hub and spoke" structure: a few hubs have huge numbers of links, while the majority of nodes have very few connections.

When studying certain types of networks, especially related to transportation and communications, we can think of them in terms of their static structure, or the dynamic processes that occur on these networks. Consider the air transportation network, for instance. We might view the picture in Figure 0.7 as a set of routes that exist between airports, independently of the actual travel that takes place on them; or as a traffic network that emerges from people moving between the airports. In the latter sense, links are diverse because they carry different amounts of traffic, and they also change over time. Both the structure and dynamics of networks

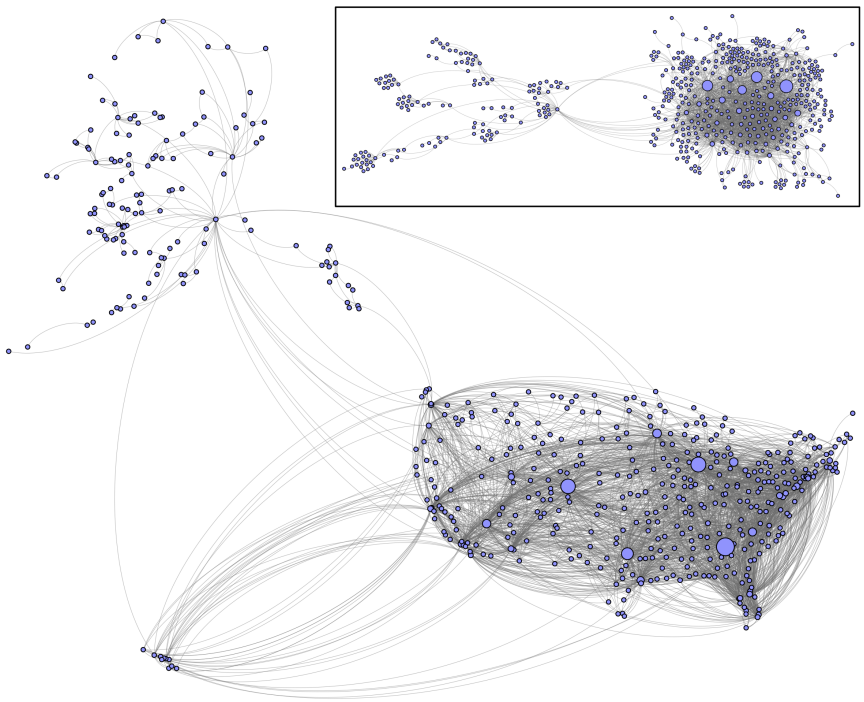

Fig. 0.7 The US air transportation network (flight data from `OpenFlights.org`). Nodes are positioned according to the geographic coordinates of the corresponding airports, so that we can make out the shape of the continental United States, Alaska, and Hawaii. Note that the map projection makes Alaska appear bigger than its actual size due to its latitude. The airport hubs with most connections (*e.g.,* Atlanta, Chicago, Denver) are clearly recognizable. The inset maps the same network, but with a different "force-directed" layout, discussed in Section 1.10.

are important. Sometimes we simply capture the dynamics by representing traffic through link directions and weights, as we discuss in Chapter 4. Other times we may wish to study the actual processes that allow a network to grow and change over time, or the interactions that take place on a network. Chapters 5 and 7 are dedicated to these topics about network dynamics.

0.6 Biological Networks

Within the cells inside our bodies, special molecules called proteins interact in a variety of ways. For example, when a protein folds, its change in structure can

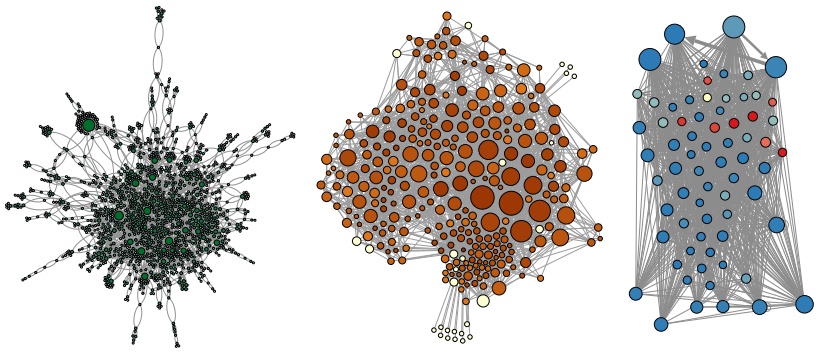

Fig. 0.8 Three biological networks. Left: protein interaction network of yeast. Node size is proportional to the number of interacting proteins. Center: neural network of the roundworm c. elegans. Large and red nodes represent neurons with more outgoing and incoming synapses, respectively. Right: Food web of species in the Florida Everglades. A directed link goes from a prey to a predator species. The weight (width) of a link represents the energy flux between the two species. Node size and color represent incoming and outgoing links, respectively, so that large blue nodes are the species at the top of food chain, while small red nodes are the species at the bottom.

regulate the function of another protein or the activity of an enzyme. Enzymes (themselves proteins) catalyze biochemical reactions and are vital to metabolism, which maintains life by harvesting energy for building and supporting the proteins that make up our tissues and organs. Proteins also regulate cell signaling and immune responses. All of these interactions can be seen as networks: protein interaction networks, metabolic networks, gene regulatory networks, and so on. These biological networks exist within a cell. At a higher level, within a body, connections between neural cells (synapses) give rise to the neural networks that form our brains. And at an even higher level, entire species interact. An animal of one species may see another species as food, creating an ecological network, or food web among species. When we think of this network, ecological balance depends on the availability of species that sustain each other. Removing a node in such a food web — when a species goes extinct, for example — affects the survival of other parts of the ecosystem network. Figure 0.8 illustrates three types of biological networks: a protein interaction network, a neural network, and a food web. They are all essential elements of life on our planet.

0.7 Summary

Networks are a general way to model and study complex systems with many interacting elements. We have seen several examples of networks. Nodes can rep-

resent many different types of objects from people to Web pages, from proteins to species, from Internet routers to airports. Nodes can have features associated with them beside labels: geographic location, wealth, activity, number of connections, and so on. Links also can represent many different kinds of relationships, from physical to virtual, from chemical to social, from communicative to informative. They can have a direction (like Web hyperlinks and email) or be reciprocal (like marriage). They can all be the same or have different features such as similarity, distance, traffic, volume, weight, and so on.

0.8 Further Readings

The use of networks to graphically represent social relationships among individuals was introduced by Moreno and Jennings (1934), who called these social networks *sociograms*.

Much more recently, studies have shown that online social networks can reveal a person's sexual orientation (Jernigan and Mistree, 2009) and facilitate highly effective phishing attacks (Jagatic et al., 2007). Conover et al. (2011b) showed that political information diffusion networks on Twitter are very polarized and segregated. As a result we can predict the political leaning of most users with high accuracy by starting with a few node labels and propagating them through network neighbors (Conover et al., 2011a).

You can read about the vision, design, and history of the Web in a book coauthored by its inventor (Berners-Lee and Fischetti, 2000).

Spring et al. (2002) explain how probes are used to measure the topology of the Internet. Achlioptas et al. (2009) show that these approaches have sampling bias. Computer scientists analyze the structure of routers and autonomous system networks to develop models called "topology generators," which can help in the design of these networks (Rossi et al., 2013). To learn more about Internet networks we recommend the book by Pastor-Satorras and Vespignani (2007).

Data about the yeast protein interaction network is from Jeong et al. (2001). C. elegans neural network data is by White et al. (1986). To learn about the human brain network, or "connectome," we recommend Sporns (2012). The Everglades ecological network is derived from Ulanowicz and DeAngelis (1998). To learn more about food webs we refer to Dunne et al. (2002); Melián and Bascompte (2004).

Data for several of the real-world network examples shown in this book is provided by the Network Repository (Rossi and Ahmed, 2015). The visualizations are done using Gephi (Bastian et al., 2009). Layout algorithms are discussed in Chapter 1.

Exercises

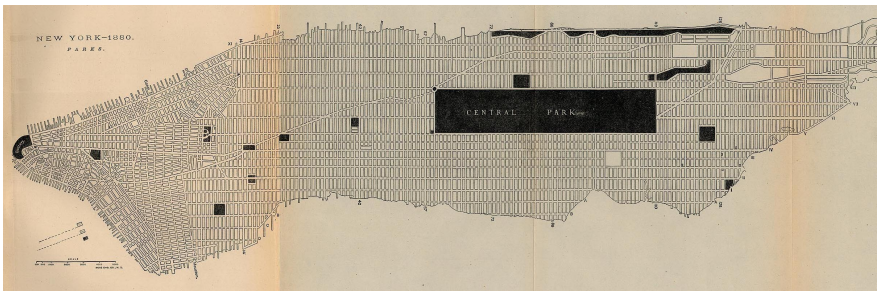

Fig. 0.9 Map of New York in 1880. From Report on the Social Statistics of Cities, Compiled by George E. Waring, Jr., United States Census Office, 1886. Image courtesy of University of Texas Libraries.

0.1 Consider the road map in Figure 0.9. If one were creating a network representation of traffic patterns, which of the following would be the best choice to make up the links of the network? (Hint: your answer to the next question may inform your answer to this question, and vice-versa.)

a. Pedestrians traveling along the streets
b. Road segments, *e.g.,* 5th Ave. between 12th and 13th streets
c. Entire roads, *e.g.,* 5th Ave.
d. Vehicles traveling on the roads

0.2 Consider the road map in Figure 0.9. In a network representation of traffic patterns, which of the following would be the best choice to make up the nodes of the network? (Hint: your answer to the previous question may inform your answer to this question, and vice-versa.)

a. City blocks, *e.g.,* the block between 5th-6th avenues and 12th-13th streets
b. Street intersections, *e.g.,* 5th Ave. and 12th St.
c. Pedestrians moving along the streets
d. Vehicles traveling on the roads

0.3 Consider the US air transportation network shown in Figure 0.7. Nodes in this network represent airports. What could a link between two airports represent?

0.4 Compare the US air transportation network in Figure 0.7 with the Manhattan road map in Figure 0.9. The air transportation network displays a distinguishing feature that the Manhattan road network lacks. What is this key characteristic?

a. Singleton nodes with no links

b. Multiple routes between nodes
c. Nodes with more than one connected link
d. Hub nodes with many links

0.5 In a social graph from Facebook, which type of link best represents the "friend" relation? Directed or undirected?

0.6 In a social graph from Twitter, which type of link best represents the "follower" relation? Directed or undirected?