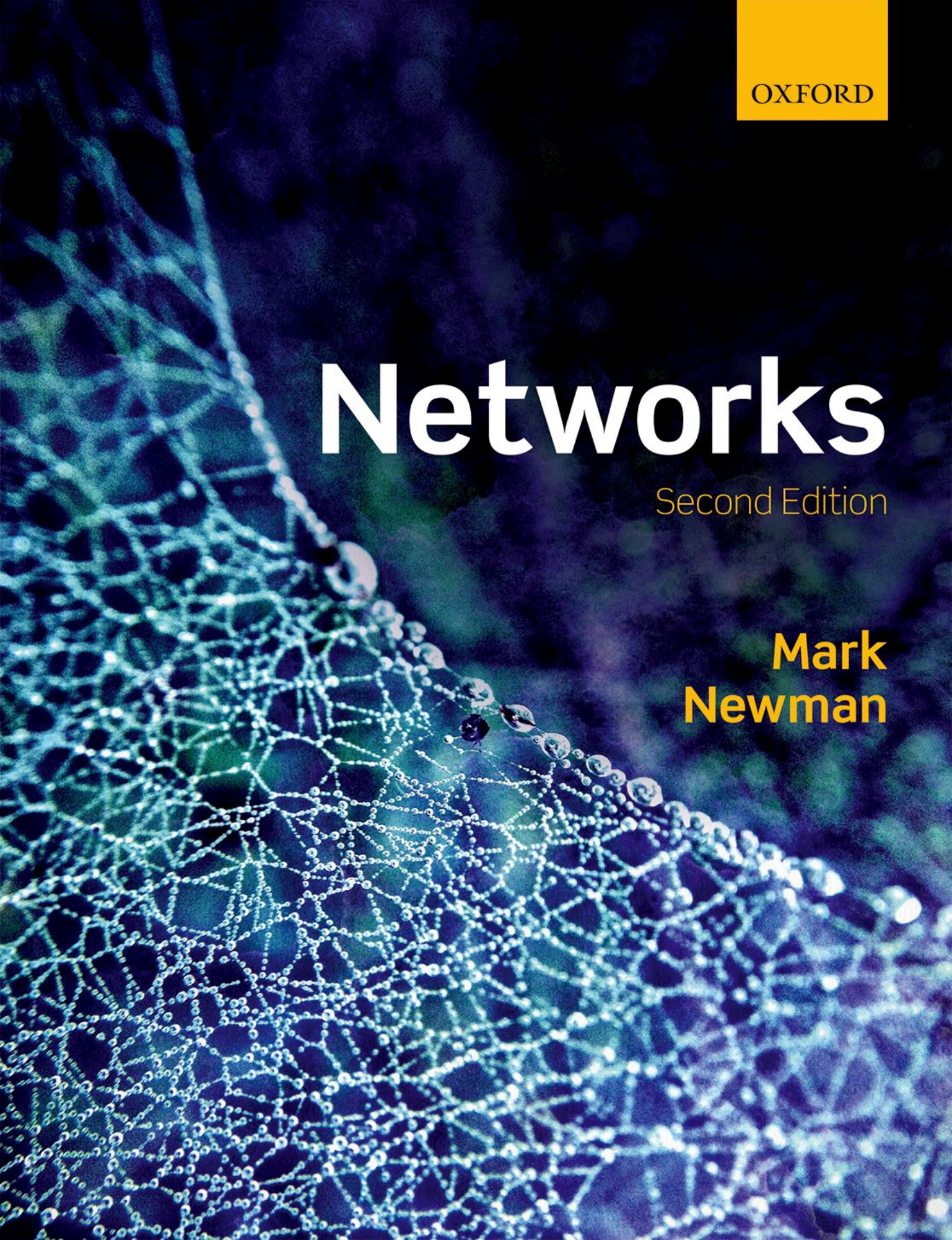

OXFORD
Networks
Second Edition
Mark
Newman

# NETWORKS

# Networks

Second Edition

Mark Newman

*University of Michigan*

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries



First Edition published in 2010

Second Edition published in 2018

Impression: 1



Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2018930384

ISBN 978–0–19–880509–0

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

DOI: 10.1093/oso/9780198805090.001.0001

# Contents

# Preface

The scientific study of networks, such as computer networks, biological networks, and social networks, is an interdisciplinary field that combines ideas from mathematics, physics, biology, computer science, statistics, the social sciences, and many other areas. The field has benefited enormously from the wide range of viewpoints brought to it by practitioners from so many different disciplines, but it has also suffered because human knowledge about networks is dispersed across the scientific community and researchers in one area often do not have ready access to discoveries made in another. The goal of this book is to bring our knowledge of networks together and present it in consistent language and notation, so that it becomes a coherent whole whose elements complement one another and in combination teach us more than any single element can alone.

The book is divided into four parts. Following a short introductory chapter, Part I describes the basic types of networks studied by present-day science and the empirical techniques used to determine their structure. Part II introduces the fundamental tools used in the study of networks, including the mathematical methods used to represent network structure, measures and statistics for quantifying network structure, and computer algorithms for calculating those measures and statistics. Part III describes mathematical models of network structure that can help us predict the behavior of networked systems and understand their formation and growth. And Part IV describes applications of network theory, including models of network resilience, epidemics taking place on networks, and network search processes.

The technical level of the presentation varies among the parts, Part I requiring virtually no mathematical knowledge for its comprehension, while Part II requires a grasp of linear algebra and calculus at the undergraduate level. Parts III and IV are mathematically more advanced and suitable for advanced undergraduates, postgraduates, and researchers working in the field. The book could thus be used as the basis of a taught course at various levels. A less technical course suitable for those with moderate mathematical knowledge might cover the material of Chapters 1 to 10, while a more technical course for

advanced students might cover the material of Chapters 6 to 13 and selected material thereafter. Each chapter from Part II onwards is accompanied by a selection of exercises that can be used to test the reader's understanding of the material.

The study of networks is a rapidly advancing field and this second edition of the book includes a significant amount of new material, including sections on multilayer networks, network statistics, community detection, complex contagion, and network synchronization. The entire book has been thoroughly updated to reflect recent developments in the field and many new exercises have been added throughout.

Over its two editions this book has been some years in the making and many people have helped me with it during that time. I must thank my ever-patient editor Sonke Adlung, with whom I have worked on various book projects for more than 25 years, and whose constant encouragement and wise advice have made working with him and Oxford University Press a real pleasure. Thanks are also due to Melanie Johnstone, Viki Kapur, Charles Lauder, Alison Lees, Emma Lonie, April Warman, and Ania Wronski for their help with the final stages of bringing the book to print.

I have benefited greatly during the writing of the book from the conversation, comments, suggestions, and encouragement of many colleagues and friends. They are, sadly, too numerous to mention exhaustively, but special thanks must go to Edoardo Airoldi, Robert Axelrod, Steve Borgatti, Elizabeth Bruch, Duncan Callaway, François Caron, Aaron Clauset, Robert Deegan, Jennifer Dunne, Betsy Foxman, Linton Freeman, Michelle Girvan, Mark Handcock, Petter Holme, Jon Kleinberg, Alden Klovdahl, Liza Levina, Lauren Meyers, Cris Moore, Lou Pecora, Mason Porter, Sidney Redner, Gesine Reinert, Martin Rosvall, Cosma Shalizi, Steve Strogatz, Duncan Watts, Doug White, Lenka Zdeborová, and Bob Ziff, as well as to the many students and other readers whose feedback helped iron out a lot of rough spots, particularly Michelle Adan, Alejandro Balbin, Ken Brown, George Cantwell, Judson Caskey, Rachel Chen, Chris Fink, Massimo Franceschet, Milton Friesen, Michael Gastner, Martin Gould, Timothy Griffin, Ruthi Hortsch, Shi Xiang Lam, Xiaoning Qian, Harry Richman, Puck Rombach, Tyler Rush, Snehal Shekatkar, Weijing Tang, Robb Thomas, Jane Wang, Paul Wellin, Daniel Wilcox, Yongsoo Yang, and Dong Zhou. I would also especially like to thank Brian Karrer, who read the entire book in draft form and gave me many pages of thoughtful and thought-provoking comments, as well as spotting a number of mistakes and typos. Responsibility for any remaining mistakes in the book of course rests entirely

with myself, and I welcome corrections from readers.

Finally, my heartfelt thanks go to my wife Carrie for her continual encouragement and support during the writing of this book. Without her the book would still have been written but I would have smiled a lot less.

Mark Newman
Ann Arbor, Michigan
June 12, 2018

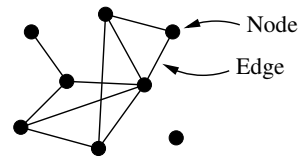
# Chapter 1

# Introduction

*A short introduction to networks*
*and why we study them*

A network is, in its simplest form, a collection of points joined together in pairs by lines. In the nomenclature of the field a point is referred to as a *node* or *vertex*[1] and a line is referred to as an *edge*. Many systems of interest in the physical, biological, and social sciences can be thought of as networks and, as this book aims to show, thinking of them in this way can lead to new and useful insights.

We begin in this first chapter with a brief introduction to some of the most commonly studied types of networks and their properties. All the topics in this chapter are covered in greater depth later in the book.




A small network composed of eight nodes and ten edges.

## Examples of networks

Networks of one kind or another crop up in almost every branch of science and technology. We will encounter a huge array of interesting examples in this book. Purely for organizational purposes, we will divide them into four broad categories: technological networks, information networks, social networks, and biological networks.

A good example of a technological network is the Internet, the computer data network in which the nodes are computers and the edges are data connections between them, such as optical fiber cables or telephone lines. Figure 1.1

[1] Plural: vertices.

*Networks*, 2nd edition. Mark Newman, Oxford University Press (2018). © Mark Newman.
DOI: 10.1093/oso/9780198805090.001.0001

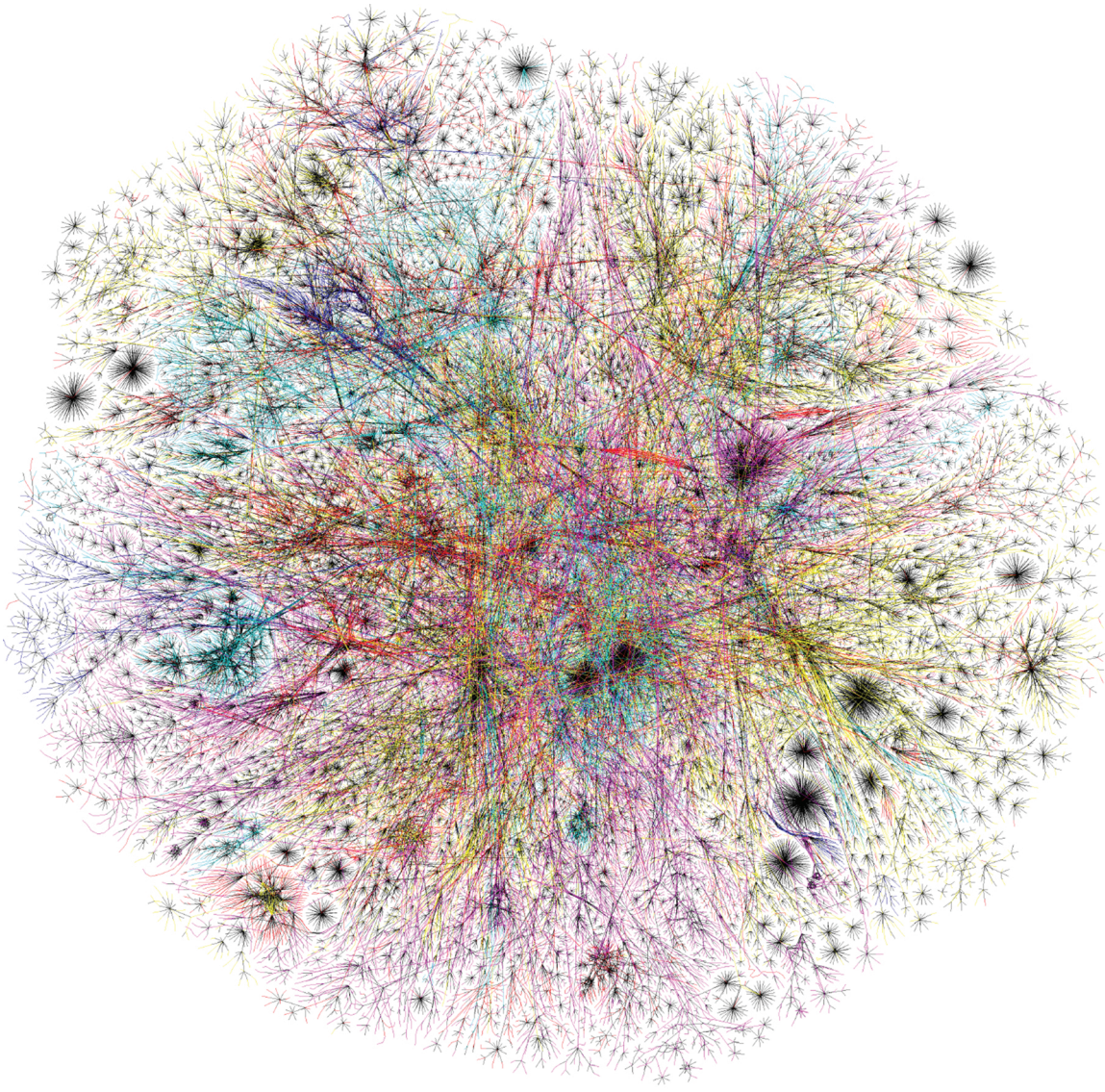


**Figure 1.1: The network structure of the Internet.** The nodes in this representation of the Internet are "class C subnets"—groups of computers with similar Internet addresses that are usually under the management of a single organization—and the connections between them represent the routes taken by Internet data packets as they hop between subnets. The geometric positions of the nodes in the picture have no special meaning; they are chosen simply to give a pleasing layout and are not related, for instance, to geographic position of the nodes. The structure of the Internet is discussed in detail in Section 2.1. Figure created by the Opte Project (`http://www.opte.org`). Reproduced with permission.

shows a picture of the structure of the Internet, a snapshot of the network as it was in 2003, reconstructed by observing the paths taken across the network by a large number of Internet data packets. It is a curious fact that although the Internet is a man-made and carefully engineered network, we don't know exactly what its structure is because it was built by many different groups of people with only limited knowledge of each other's actions and little centralized control. Our best current data on its structure are therefore derived from experimental measurements, such as those that produced this figure, rather than from any centrally held map or repository of knowledge.

We look at the Internet in more detail in Section 2.1.

There are a number of practical reasons why we might want to study the network structure of the Internet. The function of the Internet is to transport data between computers (and other devices) in different parts of the world, which it does by dividing the data into separate packets and shipping them from node to node across the network until they reach their intended destination. The network structure of the Internet will affect how efficiently it performs this function, and if we know that structure we can address many questions of practical relevance. How should we choose the route by which data are transported? Is the shortest route, geographically speaking, always necessarily the fastest? If not, then what is, and how can we find it? How can we avoid bottlenecks in the traffic flow that might slow things down? What happens when a node or an edge fails (which they do with some regularity)? How can we devise schemes to route around such failures? If we have the opportunity to add new capacity to the network, where should it be added?

Other examples of technological networks include the telephone network, networks of roads, rail lines, or airline routes, and distribution networks such as the electricity grid, water lines, oil or gas pipelines, or sewerage pipes. Each of these networks raises questions of their own: what is their structure, how does it affect the function of the system, and how can we design or change the structure to optimize performance? In some cases, such as airline routes, networks are already highly optimized; in others, such as the road network, the structure may be largely a historical accident and is in some cases far from optimal.

Our second class of networks are the information networks, a more abstract class that represents the network structure of bodies of information. The classic example is the World Wide Web. We discussed the Internet above, but the Web is not the same thing as the Internet, even though the two words are often used interchangeably in casual speech. The Internet is a physical network of computers linked by actual cables (or sometimes radio links) running between them. The Web, on the other hand, is a network of web pages and the links between them. The nodes of the World Wide Web are the web pages and the

Information networks are discussed at length in Chapter 3.

The World Wide Web is discussed in more detail in Section 3.1.

edges are "hyperlinks," the highlighted snippets of text or push-buttons on web pages that we click on to navigate from one page to another. A hyperlink is purely a software construct; you can link from your web page to a page that lives on a computer on the other side of the world just as easily as you can to a friend down the hall. There is no physical structure, like an optical fiber, that needs to be built when you make a new link. The link is merely an address that tells the computer where to look next when you click on it. Thus the network structure of the Web and the Internet are completely distinct.

Abstract though it may be, the World Wide Web, with its billions of pages and links, has proved enormously useful, not to mention profitable, and the structure of the network is of substantial interest. Since people tend to add hyperlinks between pages with related content, the link structure of the Web reveals something about relationships between content and topics. Arguably, the structure of the Web could be said to reflect the structure of human knowledge. What's more, people tend to link more often to pages they find useful than to those they do not, so that the number of links pointing to a page can be used as a measure of its usefulness. A more sophisticated version of this idea lies behind the operation of the popular web search engine *Google*, as well as some others.

The mechanics of web search are discussed in Section 18.1.

The Web also illustrates another concept of network theory, the *directed network*. Hyperlinks on the Web run in one specific direction, from one web page to another. You may be able to click a link on page A and get to page B, but there is no requirement that B has a link back to A again. (It might contain such a link but it doesn't have to.) One says that the edges in the World Wide Web are *directed*, running from the linking page to the linked.

Another much-studied example of an information network is a citation network, such as the network of citations between academic journal articles. Academic articles typically include a bibliography of references to other previously published articles, and one can think of these references as forming a network in which the articles are the nodes and there is a directed edge from article A to article B if A cites B in its bibliography. As with the World Wide Web, one can argue that such a network reflects, at least partially, the structure of the body of knowledge contained in the articles, with citations between articles presumably indicating related content. Indeed there are many similarities between the Web and citation networks and a number of the techniques developed for understanding and searching the Web have in recent years started to be applied to citation networks too, to help scientists and others filter the vast amount of published research and data to find useful papers.

Our third broad class of networks are the social networks. When one talks about "social networks" today, most of us think of online services such as

*Facebook* or *Twitter*, but in the scientific literature the term is used much more broadly to encompass any network in which the nodes are people (or sometimes groups of people, such as firms or teams) and the edges between them are social connections of some kind, such as friendship, communication, or collaboration. The field of sociology has perhaps the longest and best developed tradition of the empirical study of networks as they occur in the real world, and many of the mathematical and statistical tools used in the study of networks are borrowed, directly or indirectly, from sociologists.

Social networks are discussed in more depth in Chapter 4.

Figure 1.2 shows a famous example of a social network from the sociology literature, Wayne Zachary's "karate club" network. This network represents the pattern of friendships among the members of a karate club at a North American university, reconstructed from observations of social interactions between them. Sociologists have performed a huge number of similar studies over the decades, including studies of friendship patterns among CEOs of corporations, doctors, monks, students, and conference participants, and networks of who works with whom, who does business with whom, who seeks advice from whom, who socializes with whom, and who sleeps with whom. Such studies, in which data are typically collected by hand, are quite arduous, so the networks they produce are usually small, like the one in Fig. 1.2, which has just 34 nodes. But in recent years, much larger social networks have been assembled using, for instance, online data from Facebook and similar services. At the time of writing, Facebook had over two billion users worldwide—more than a quarter of the population of the world—and information on the connection patterns between all of them. Many online social networking companies, including Facebook, have research divisions that collaborate with the academic community to do research on social networks using their vast data resources.

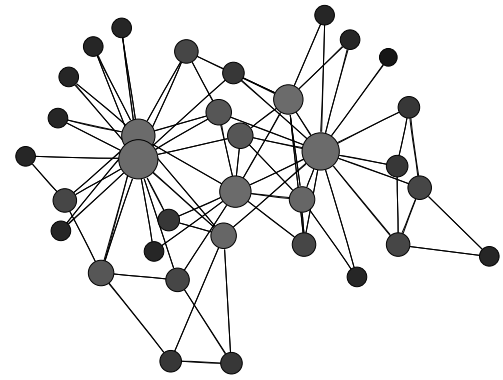


**Figure 1.2: Friendship network between members of a club.** This social network from a study conducted in the 1970s shows the pattern of friendships between the members of a karate club at an American university. The data were collected and published by Zachary [479].

Our fourth and final class of networks is biological networks. Networks occur in range of different settings in biology. Some are physical networks like neural networks—the connections between neurons in the brain—while others are more abstract. In Fig. 1.3 we show a picture of a "food web," an ecological network in which the nodes are species in an ecosystem and the edges represent predator–prey relationships between them. That is, a pair of species is connected by an edge in this network if one species eats the other. The study of food webs can help us understand and quantify many ecological phenomena, particularly concerning energy and carbon flows and the interdependencies

Neural networks are discussed in Section 5.2 and food webs in Section 5.3.

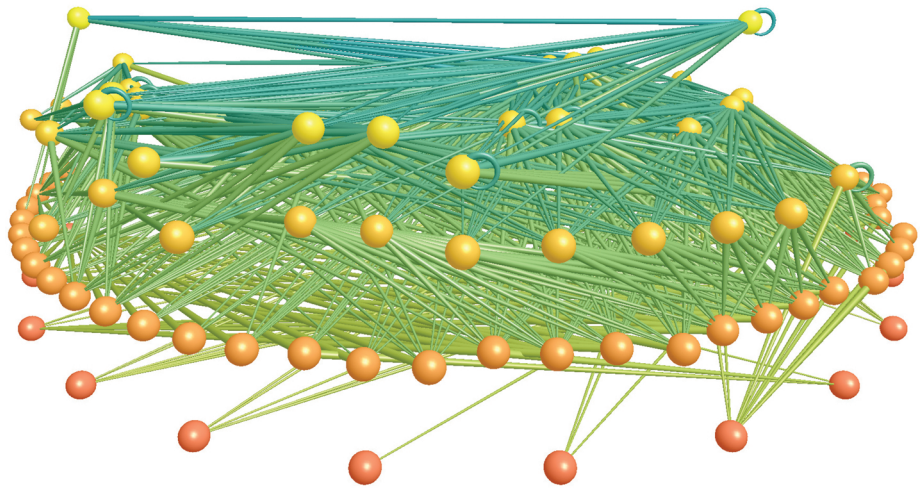


**Figure 1.3: The food web of Little Rock Lake, Wisconsin.** This elegant picture summarizes the known predatory interactions between species in a freshwater lake in the northern United States. The nodes represent the species and the edges run between predator–prey species pairs. The vertical position of the nodes represents, roughly speaking, the trophic level of the corresponding species. The figure was created by Richard Williams and Neo Martinez [321].

between species. Food webs also provide us with another example of a directed network, like the World Wide Web and citation networks discussed previously. If species A eats species B then probably B does not also eat A, so the relationship between the two is a directed one.

Biochemical networks are discussed in detail in Section 5.1.

Another class of biological networks are the biochemical networks. These include metabolic networks, protein–protein interaction networks, and genetic regulatory networks. A metabolic network, for instance, is a representation of the pattern of chemical reactions that fuel the cells in an organism. The reader may have seen the wallcharts of metabolic reactions that adorn the offices of some biochemists, incredibly detailed maps with hundreds of tiny inscriptions linked by a maze of arrows. The inscriptions—the nodes in this network—are metabolites, the chemicals involved in metabolism, and the arrows—directed edges—are reactions that turn one metabolite into another. The representation of reactions as a network is one of the first steps towards making sense of the bewildering array of biochemical data generated by current experiments in biochemistry and molecular genetics.

An example metabolic network map appears as Fig. 5.2 on page 75.

These are just a few examples of the types of networks that will concern us

in this book. These and many others are studied in more detail in the following chapters.

### WHAT CAN WE LEARN FROM NETWORKS?

Networks capture the pattern of interactions between the parts of a system. It should come as no surprise (although in some fields it is a relatively recent realization) that the pattern of interactions can have a big effect on the behavior of a system. The pattern of connections between computers on the Internet, for instance, affects the routes that data take over the network and hence the efficiency with which the network transports those data. The connections in a friendship network affect how people learn, form opinions, and gather news, as well as other less obvious phenomena, such as the spread of disease. Unless we know something about the structure of these networks, we cannot hope to understand fully how the corresponding systems work.

A network is a simplified representation that reduces a system to an abstract structure or *topology*, capturing only the basics of connection patterns and little else. The systems studied can, and often do, have many other interesting features not represented by the network—the detailed behaviors of individual nodes, such as computers or people, for instance, or the precise nature of the interactions between them. Some of these subtleties can be captured by embroidering the network with labels on the nodes or edges, such as names or strengths of interactions, but even so a lot of information is usually lost in the process of reducing a full system to a network representation. This has some disadvantages but it has advantages as well.

Some common network extensions and variants are discussed in Chapter 6.

Scientists in a wide variety of fields have, over the years, developed an extensive set of mathematical and computational tools for analyzing, modeling, and understanding networks. Some of these tools start from a simple network topology—a set of nodes and edges—and after some calculation tell you something potentially useful about the network: which is the best connected node, say, or how similar two nodes are to one another. Other tools take the form of network models that can make mathematical predictions about processes taking place on networks, such as the way traffic will flow over the Internet or the way a disease will spread through a community. Because they work with networks in their abstract form, tools such as these can be applied to almost any system that has a network representation. Thus, if there is a system you are interested in, and it can usefully be represented as a network, then there are hundreds of ready-made tools out there, already fully developed and well understood, that you can immediately apply to your system. Not all of them will necessarily give useful results—which measurements or calculations are

useful for a particular system depends on what the system is and does and on what specific questions you are trying to answer about it. Still, if you have a well-posed question about a networked system there will, in many cases, already be a tool available that will help you address it.

Networks are thus a general means for representing the structure of a system that creates a bridge between empirical data and a large toolkit of powerful analysis techniques. In this book we discuss many examples of specific networks in different fields, along with techniques for their analysis drawn from mathematics, physics, the computer and information sciences, the social sciences, biology, and elsewhere. In doing so, we will bring together a wide range of ideas and expertise from many disciplines to build a comprehensive understanding of the science of networks.

## PROPERTIES OF NETWORKS

Perhaps the most fundamental question we can ask about networks is this: if we know the shape of a network, what can we learn about the nature and function of the system it describes? In other words, how are the structural features of a network related to the practical issues we care about? This question is essentially the topic of this entire book, and we are not going to answer it in this chapter alone. Let us, however, look briefly here at a few representative concepts, to get a feel for the kinds of ideas we will be dealing with.

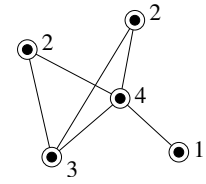
A first step in analyzing the structure of a network is often to make a picture of it. Figures 1.1, 1.2, and 1.3 are typical examples. Each of them was generated by a specialized computer program designed for network visualization and there are many such programs available, both commercially and for free, if you want to produce pictures like these for yourself. Visualization can be an extraordinarily useful tool in the analysis of network data, allowing one to instantly see important structural features that would otherwise be difficult to pick out of the raw data. The human eye is enormously gifted at discerning patterns, and visualizations allow us to put this gift to work on our network problems.

On the other hand, direct visualization of networks is only really useful for networks up to a few hundreds or thousands of nodes, and for networks that are relatively sparse, meaning that the number of edges is quite small. If there are too many nodes or edges then pictures of the network will be too complicated for the eye to comprehend and their usefulness becomes limited. Many of the networks that scientists are interested in today have hundreds of thousands or even millions of nodes, which means that visualization is not of much help and we need to employ other techniques to understand them. Moreover, while the

eye is definitely a powerful tool for data analysis, it is not a wholly reliable one, sometimes failing to pick out important patterns in data or even seeing patterns where they don't exist. To address these issues, network theory has developed a large toolchest of measures and metrics that can help us understand what networks are telling us, even in cases where useful visualization is impossible or unreliable.

An example of a useful (and widely used) class of network metrics are the *centrality* measures. Centrality quantifies how important nodes are in a network, and social network analysts in particular have expended considerable effort studying it. There are, of course, many different possible concepts or definitions of what it means for a node to be central in a network, and there are correspondingly many centrality measures. Perhaps the simplest of them is the measure called *degree*. The degree of a node in a network is the number of edges attached to it. In a social network of friendships, for instance, such as the network of Fig. 1.2, the degree of an individual is the number of friends he or she has within the network. For the Internet degree would be the number of data connections a computer has. In many cases the nodes with the highest degrees in a network, those with the most connections, also play major roles in the functioning of the system, and hence degree can be a useful guide for focusing our attention on the system's most important elements.

See Chapter 7 for further discussion of centrality measures.




The number beside each node in this small network indicates the node's degree.

In undirected networks degree is just a single number, but in directed networks nodes have two different degrees, *in-degree* and *out-degree*, corresponding to the number of edges pointing inward and outward respectively. For example, the in-degree of a web page is the number of other pages that link to it, while the out-degree is the number of pages to which it links. We have already mentioned one example of how centrality can be put to use on the Web to answer an important practical question: by counting the number of links a web page gets—the in-degree of the page—a search engine can identify pages that are likely to contain useful information.

A further observation concerning degree is that many networks are found to contain a small but significant number of "hubs"—nodes of unusually high degree. Social networks, for instance, often contain a few individuals with an unusually large number of acquaintances. The Web has a small fraction of websites with a very large number of links. There are a few metabolites that take part in a very large number of metabolic processes. A major topic of research in recent years has been the investigation of the effects of hubs on the performance and behavior of networked systems. A wide range of results, both empirical and theoretical, indicate that hubs can have a disproportionate effect, particularly on network resilience and transport phenomena, despite being few in number.

Hubs are discussed further in Section 10.3.

The small-world effect is discussed further in Sections 4.6 and 10.2.

Another example of a network concept that arises repeatedly and has real practical implications is the so-called *small-world effect*. Given a network, one can ask what the shortest distance is, through the network, between a given pair of nodes. In other words, what is the minimum number of edges one would have to traverse in order to get from one node to the other? For instance, your immediate friend would have distance 1 from you in a network of friendships, while a friend of a friend would have distance 2. It has been found empirically (and can be proven mathematically in some cases) that the mean distance between node pairs in many networks is very short, often no more than a dozen steps or so, even for networks with millions of nodes or more. Although first studied in the context of friendship networks, this small-world effect appears to be widespread, occurring in essentially all types of networks. In popular culture it is referred to as the "six degrees of separation," after a successful stage play and film of the same name in which the effect is discussed. The (semi-mythological) claim is that you can get from anyone in the world to anyone else via a sequence of no more than five intermediate acquaintances—six steps in all.

The small-world effect has substantial repercussions. For example, news and gossip spread over social networks—if you hear an interesting rumor from a friend, you may pass it on to your other friends, and they in turn may pass it on to theirs, and so forth. Clearly a rumor will spread faster and further if it only takes six steps to reach anyone in the world than if it takes a hundred, or a million. And indeed it is a matter of common experience that a suitably scandalous rumor can reach the ears of an entire community in what seems like the blink of an eye.

Or consider the Internet. One of the reasons the Internet functions at all is because any computer on the network is only a few hops across the network from any other. Typical routes taken by data packets over the Internet rarely have more than about twenty hops, and certainly the performance of the network would be much worse if packets had to make a thousand hops instead. In effect, our ability to receive data near instantaneously from anywhere in the world is a direct consequence of the small-world effect.

Community structure in networks is discussed in detail in Chapter 14.

A third example of a network phenomenon of practical importance is the occurrence of clusters or communities in networks. We are most of us familiar with the idea that social networks break up into subcommunities. In friendship networks, for instance, one commonly observes groups of close friends within the larger, looser network of passing acquaintances. Similar clusters occur in other types of network as well. The Web contains clusters of web pages that all link to one another, perhaps because they are about the same topic, or they all belong to the same company. Metabolic networks contain groups of metabolites

that interact with one another to perform certain biochemical tasks. And if it is the case that clusters or groups correspond to functional divisions in this way, then we may be able to learn something by taking a network and decomposing it into its constituent clusters. The way a network breaks apart can reveal levels and concepts of organization that are not easy to see by other means.

The detection and analysis of clusters in networks is an active topic at the frontier of current networks research, holding promise for exciting applications in the future.

### OUTLINE OF THIS BOOK

This book is divided into four parts. In the first part, consisting of Chapters 2 to 5, we introduce the various types of network encountered in the real world, including technological, social, and biological networks, and the empirical techniques used to discover their structure. Although it is not the purpose of this book to describe any one particular network in great detail, the study of networks is nonetheless firmly founded on empirical observations and a good understanding of what data are available and how they are obtained is immensely helpful in understanding the science of networks as it is practiced today.

The second part of the book, Chapters 6 to 10, introduces the fundamental theoretical ideas and methods on which our current understanding of networks is based. Chapter 6 describes the basic mathematics used to capture network ideas, while Chapter 7 describes the measures and metrics we use to quantify network structure. Chapter 8 describes the computer methods that are crucial to practical calculations on today's large networks, Chapter 9 describes methods of network statistics and the role of errors and uncertainty in network studies, and Chapter 10 describes some of the intriguing patterns and principles that emerge when we apply all of these ideas to real-world network data.

In the third part of the book, Chapters 11 to 13, we look at mathematical models of networks, including both traditional models, such as random graphs and their extensions, and newer models, such as models of growing networks and community structure. The material in these chapters forms a central part of the canon of the field and has been the subject of a vast amount of published scientific research.

Finally, in the fourth and last part of the book, Chapters 14 to 18, we look at applications of network theory to a range of practical questions, including community detection, network epidemiology, dynamical systems, and network search processes. Research is less far advanced on these topics than it is in other areas of network science and there is much we do not know. The final chapters

of the book probably raise at least as many questions as they answer, but this, surely, is a good thing. For those who would like to get involved, there are plenty of fascinating open problems waiting to be addressed.

# Part I

# The empirical study of networks

# Chapter 2

# Technological networks

*A discussion of engineered networks like the Internet and the power grid and methods for determining their structure*

In the next four chapters we describe and discuss some of the most commonly studied networks, dividing them into four broad classes—technological networks, information networks, social networks, and biological networks. For each class we list some important examples and examine the techniques used to measure their structure.

The four classes are not rigorously defined and there is, as we will see, some overlap between them, with some networks plausibly belonging to two or more classes. Nonetheless, the division into classes is a useful one, since networks in the same class are often treated using similar techniques or ideas.

It is not our intention in this book to study any one network in great detail. Plenty of other books do that. Nonetheless, network science is concerned with understanding and modeling the behavior of real-world systems and observational data are the starting point for essentially all the developments of the field, so it will be useful to have a grasp of the types of networks commonly studied and the data that describe them. In this chapter we look at technological networks, the physical infrastructure networks that form the backbone of modern technological societies. Perhaps the most celebrated such network—and a relatively recent entry in the field—is the Internet, the global network of data connections that links computers and other information systems together. Section 2.1 is devoted to a discussion of the Internet. A number of other important examples of technological networks, including power grids, transportation networks, delivery and distribution networks, and telephone networks, are discussed in subsequent sections.

---

*Networks*, 2nd edition. Mark Newman, Oxford University Press (2018). © Mark Newman.
DOI: 10.1093/oso/9780198805090.001.0001

## 2.1 The Internet

The Internet is the worldwide network of physical data connections between computers, phones, tablets, and other devices. The Internet is a *packet-switched* data network, meaning that messages sent over it are broken up into *packets*, small chunks of data, that are sent separately over the network and reassembled into a complete message again at the other end. The format of the packets follows a standard known as the *Internet Protocol* (IP) and includes an *IP address* in each packet that specifies the packet's destination, so that it can be routed correctly across the network.

The Internet should not be confused with the World Wide Web, a virtual network of web pages and hyperlinks, which we discuss separately in Section 3.1.

The simplest network representation of the Internet (there are others, which we will discuss shortly) is one in which the nodes of the network represent computers and other devices, and the edges represent data connections between them, such as optical fiber lines or wireless connections. In fact, ordinary computers and other consumer devices mostly occupy the nodes on the "outside" of the network, the end points (or starting points) of data flows, and do not act as intermediate points between others. (Indeed, most end-user devices only have a single connection to the Net, so it would not be possible for them to lie on a path between any others.) The "interior" nodes of the Internet are primarily *routers*, powerful special-purpose machines at the junctions between data lines that receive data packets and forward them in one direction or another towards their intended destination (essentially larger versions of the network router you might have in your home).

The general overall shape of the Internet is shown, in schematic form, in Fig. 2.1. The network is composed of three levels or circles of nodes. The innermost circle, the core of the network, is called the *backbone* and contains the trunk lines that provide long-distance high-bandwidth data transport across the globe, along with the high-performance routers and switching centers that link the trunk lines together. The trunk lines are the highways of the Internet, built with the fastest fiber optic connections available (and improving all the time). The backbone is owned and operated by a set of *network backbone providers* (NBPs), who are primarily national governments and major telecommunications companies such as Level 3 Communications, Cogent, NTT, and others.

The second circle of the Internet is composed of *Internet service providers* or ISPs—commercial companies, governments, universities, and others who contract with NBPs for connection to the backbone and then resell or otherwise provide that connection to end users, the ultimate consumers of Internet bandwidth, who form the third circle—businesses, government offices, academic institutions, people in their homes, and so forth. As Fig. 2.1 shows, the ISPs

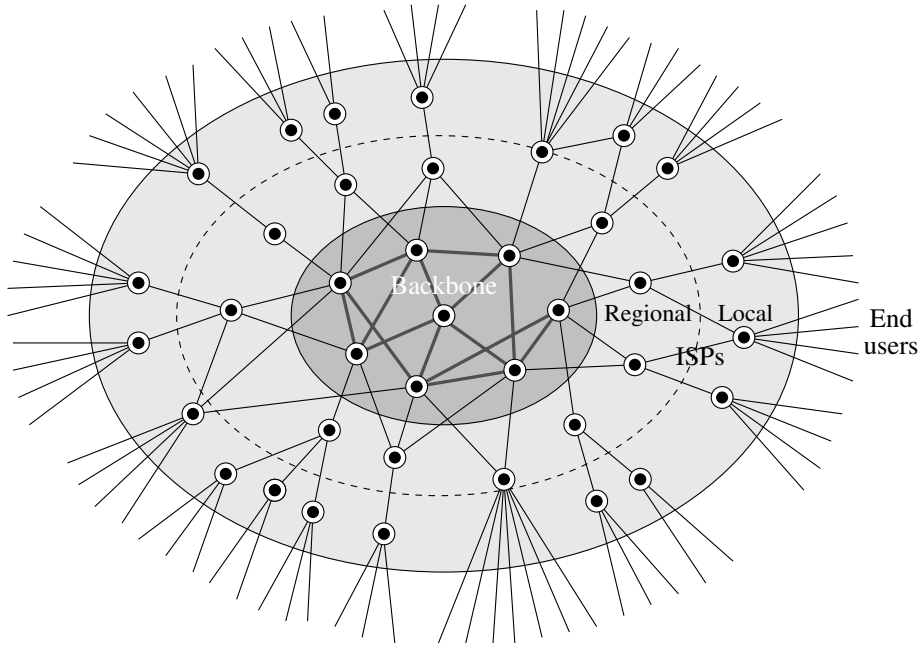



**Figure 2.1: A schematic depiction of the structure of the Internet.** The nodes and edges of the Internet fall into a number of different classes: the backbone of high-bandwidth long-distance connections; the ISPs, who connect to the backbone and who are divided roughly into regional (larger) and local (smaller) ISPs; and the end users—home users, companies, and so forth—who connect to the ISPs.

are further subdivided into *regional ISPs* and *local* or *consumer ISPs*, the former being larger organizations whose primary customers are the local ISPs, who in turn sell network connections to the end users. This distinction is somewhat blurred however, because large consumer ISPs, such as AT&T or British Telecom, often act as their own regional ISPs (and some may be backbone providers as well).

The network structure of the Internet is not dictated by any central authority. Protocols and guidelines are developed by an informal volunteer organization called the Internet Engineering Task Force, but one does not have to apply to any central Internet authority for permission to build a new spur on the network, or to take one out of service.

One of the remarkable features of the Internet is the scheme used for routing data across the network, in which the paths that packets take are determined by automated negotiation among Internet routers under a system called the

*Border Gateway Protocol* (BGP). BGP is designed in such a way that if new nodes or edges are added to the network or old ones disappear, either permanently or temporarily, routers will take note and adjust their routing policy appropriately. There is a certain amount of human oversight involved, to make sure the system keeps running smoothly, but no "Internet government" is needed to steer things from on high; the system organizes itself by the combined actions of many local and essentially autonomous computer systems.

While this is an excellent feature of the system from the point of view of robustness and flexibility, it is a problem for those who want to study the structure of the Internet. If there were a central Internet government with a complete map of the system, then the job of determining the network structure would be easy—one would just look at the map. But there is no such organization and no such map. Instead the network's structure must be determined by experimental measurements. There are two primary methods for doing this. The first uses "traceroute"; the second uses BGP.

### 2.1.1 Measuring Internet structure using traceroute

There is currently no simple means by which to probe the network structure of the Internet directly. We can, however, quite easily discover the particular path taken by data packets sent from one computer to another on the Internet. The standard tool for doing this is called *traceroute*.

Each Internet data packet contains, among other things, a destination address, which says where it is going; a source address, which says where it started from; and a *time-to-live* (TTL). The TTL is a number that specifies the maximum number of hops that the packet can make to get to its destination, a hop being the traversal of one edge in the network. At every hop, the TTL is decreased by one, and if it reaches zero the packet is discarded, meaning it is deleted and not forwarded any further over the network. A message is also then transmitted back to the sender informing them that the packet was discarded and where it got to. In this way the sender is alerted if data is lost, allowing them to resend the contents of the packet if necessary. The TTL exists mainly as a safeguard to prevent packets from losing their way on the Internet and wandering around forever, but we can make use of it to track packet progress as well. The idea is as follows.

First, we send out a packet with the destination address of the network node we are interested in and a TTL of 1. The packet makes a single hop to the first router along the way, its TTL is decreased to 0, the packet is discarded by the router, and a message is returned to us telling us, among other things, the IP address of the router. We record this address and then repeat the process

with a TTL of 2. This time the packet makes two hops before dying and the returned message tells us the IP address of the second router along the path. The process is repeated with larger and larger TTL until the destination is reached, and the set of IP addresses received as a result tells us the entire route taken to get there.[1] There are standard software tools that will perform the complete procedure automatically and print out the list of IP addresses for us. On many operating systems the tool that does this is called "traceroute."[2]

We can use traceroute (or a similar tool) to probe the network structure of the Internet. The idea is to assemble a large data set of traceroute paths between many different pairs of points on the Internet. With luck, most of the edges in the network (though usually not all of them) will appear in at least one of these paths, and the combination of all of them together should give a reasonably complete picture of the network. Early studies, for the sake of expediency, limited themselves to paths starting from just a few source computers, but more recent ones make use of distributed collections of thousands of sources to develop a very complete picture of the network.

The paths from any single source to a set of destinations form a branching structure as shown schematically in Figs. 2.2a, b, and c.[3] The source computers should, ideally, be well distributed over the network. If they are close together then there may be substantial overlap between the paths to distant nodes, meaning that they will needlessly duplicate each other's efforts rather than returning independent measurements.

Once one has a suitable set of traceroute paths, a simple union of them gives us our snapshot of the network structure—see Fig. 2.2d. That is, we create a node in our network for every unique IP address that appears at least once in any of the paths and an edge between any pair of addresses that fall on adjacent steps of any path. As hinted above, it is unlikely that this procedure

[1] We are assuming that each packet takes the same route to the destination. It is possible, though rare, for different packets to take different routes, in which case the set of IP addresses returned by the traceroute procedure will not give a correct path through the network. This can happen, for instance, if congestion patterns along the route vary significantly while the procedure is being performed, causing the network to reroute packets along less congested paths. Serious Internet mapping experiments perform repeated traceroute measurements to minimize the errors introduced by effects such as these.

[2] On the Windows operating system it is called "tracert." On some Linux systems it is called "tracepath."

[3] If there were a unique best path to every node, then the set of paths would be a "tree," meaning it would contain no loops. (See Section 6.8 for a discussion of trees.) Because of the way routing algorithms work, however, this is not always the case in practice—two routes that originate at the same point and pass through the same node on the way to their final destination can still take different routes to get to that node, so that the set of paths can contain loops.

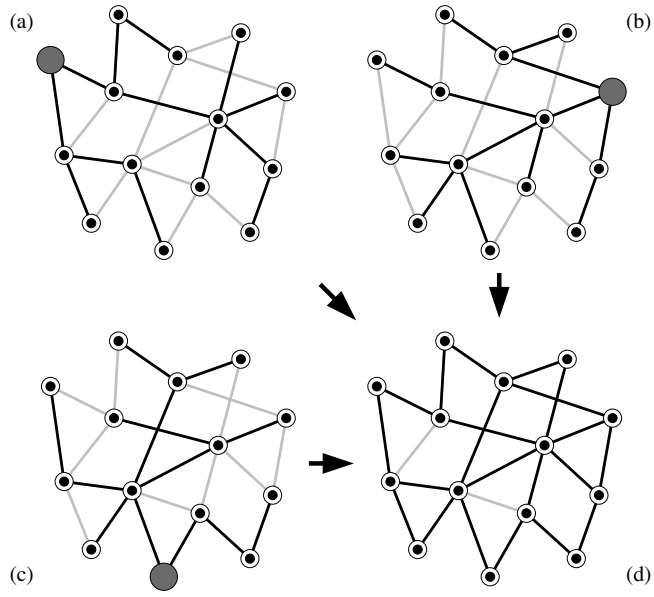



**Figure 2.2: Reconstruction of the topology of the Internet from traceroute data.** In (a), (b), and (c) we show in bold the edges that fall along traceroute paths starting from the three highlighted source nodes. In (d) we form the union of these edges to make a picture of the overall network topology. Note that a few edges are still missing from the final picture (the remaining gray edges in (d)) because they happen not to appear in any of the three individual traceroute data sets.

will find all the edges in the network (see Fig. 2.2d again), and for studies based on small numbers of sources there can be significant biases in the sampling of edges [3, 284]. However, better and better data sets are becoming available as time passes, and it is believed that we now have a reasonably complete picture of the shape of the Internet.

In fact, complete (or near-complete) representations of the Internet of the kind described here can be cumbersome to work with and are typically not used directly for network studies. There are billions of distinct IP addresses in use on the Internet at any one time, with many of those corresponding to end-user devices that appear or disappear as the devices are turned on or off or connections to the Internet are made or broken. Most studies of the Internet ignore end users and restrict themselves to just the routers, in effect concentrating on the inner zones in Fig. 2.1 and ignoring the outermost one. We will refer to such maps of the Internet as representations at the *router level*. The nodes in the network are routers, and the edges between them are network connections.

It may appear strange to ignore end-user devices, since the end users are, after all, the entire reason for the Internet's existence in the first place. However, it is the structure of the network at the router level that is responsible for most aspects of the performance, robustness, and efficiency of the network, that dictates the patterns of traffic flow on the network, and that forms the focus of most work on Internet structure and design. To the extent that these are the issues of scientific interest, therefore, it makes sense to concentrate our efforts on the router-level structure.

An example of a study of the topology of the Internet at the router level is that of Faloutsos *et al.* [168], who looked at the "degree distribution" of the network and discovered it to follow, approximately, a power law. We discuss degree distributions and power laws in networks in more detail in Section 10.4.

Even after removing all or most end users from the network, the structure of the Internet at the router level may still be too detailed for our purposes. Often we would like a more coarse-grained representation of the network that gives us a broader overall picture of network structure. Such representations can be created by grouping sets of IP addresses together into single nodes. Three different ways of grouping addresses are in common use, giving rise to three different coarse-grained representations, at the level of subnets, domains, and autonomous systems.

A *subnet* is a group of IP addresses defined as follows. IP addresses consist of four numbers, each one in the range from 0 to 255 (eight bits in binary) and typically written in a string separated by periods or dots.[4] For example, the IP address of the main web server at the author's home institution, the University of Michigan, is 141.211.243.44. IP addresses are allocated to organizations in blocks. The University of Michigan, for instance, owns (among others) all the addresses of the form 141.211.243.xxx, where "xxx" can be any number between 0 and 255. Such a block, where the first three numbers in the address are fixed and the last can be anything, is called a *class C subnet*. There are also class B subnets, which have the form 141.211.xxx.yyy, and class A subnets, which have the form 141.xxx.yyy.zzz.

Since all the addresses in a class C subnet are usually allocated to the same organization, a reasonable way of coarse-graining the Internet's network structure is to group nodes into class C subnets. In most cases this will group

[4]This description applies to addresses as they appear in IP version 4, which is the most widely used version of the protocol. A new version, version 6, which uses longer addresses, is slowly gaining acceptance, but it has a long way to go before it becomes as popular as its predecessor. (IP versions 1, 2, 3, and 5 were all experimental and were never used widely. Versions 4 and 6 are the only two that have seen widespread use.)

together nodes in the same organization, although larger organizations, like the University of Michigan, may own more than one class C subnet, so there will still be more than one node in the coarse-grained network corresponding to such organizations.

Given the topology of the network in terms of individual IP addresses, it is an easy matter to lump together into a single node all addresses in each class C subnet and place an edge between any two subnets if any address in one has a network connection to any address in the other. Figure 1.1 on page 2 shows an example of the network structure of the Internet at the level of class C subnets.

The second common type of coarse-graining is coarse-graining at the domain level. A *domain* is a group of computers and routers under, usually, the control of a single organization and identified by a single *domain name*, normally the last two or three parts of a computer's address when the address is written in human-readable text form (as opposed to the numeric IP addresses considered above). For example, "`umich.edu`" is the domain name for the University of Michigan and "`oup.com`" is the domain name for Oxford University Press. The name of the domain to which a computer belongs can be determined from the computer's IP address by a "reverse DNS lookup," a network service set up to provide precisely this type of information. Thus, given the network topology in terms of IP addresses, it is a straightforward task to determine the domain to which each IP address belongs and group nodes in the network according to their domain. Then an edge is placed between two nodes if any IP address in one has a direct network connection to any address in the other. The study by Faloutsos *et al.* [168] mentioned earlier looked at this type of domain-level structure of the Internet as well as the router-level structure.

The third common coarse-graining of the network is coarse-graining at the level of autonomous systems. This type of coarse-graining, however, is not usually used with traceroute data but with data obtained using an alternative method based on BGP routing tables, for which it forms the most natural unit of representation. The BGP method and autonomous systems are discussed in the next section.

### 2.1.2 Measuring Internet structure using routing tables

Internet routers maintain *routing tables* that allow them to decide in which direction incoming packets should be sent to best reach their destination. Routing tables are constructed from information shared between routers using BGP. They consist of lists of complete paths from the router in question to destinations on the Internet. When a packet arrives at a router, the router examines it to determine its destination and looks up that destination in the routing table.

The first step of the path in the appropriate table entry tells the router how the packet should be sent on its way.

In theory routers need store only the first step on each path in order to route packets correctly. However, for efficient calculation of routes using BGP it is highly desirable that routers be aware of the entire path to each destination, and since the earliest days of the Internet all routers have operated in this way. We can make use of this fact to measure the structure of the Internet.

Routing tables in routers are represented at the level of *autonomous systems*. An autonomous system (or AS) is a collection of routers, computers, or other devices, usually under single administrative control, within which data routing is handled independently of the wider Internet (hence the name "autonomous system"). That is, when a data packet arrives at a router belonging to an autonomous system, destined for a specific device or user within that same autonomous system, it is the responsibility of the autonomous system to get the packet the last few steps to its final destination. Data passing between autonomous systems, however, is handled by the Internet-wide mechanisms of BGP. Thus it's necessary for BGP to know about routing only down to the level of autonomous systems and hence BGP tables are most conveniently represented in autonomous system terms. In practice, autonomous systems, of which there are (at the time of writing) about fifty thousand on the Internet, often coincide with domains, or nearly so.

Autonomous systems are assigned unique identification numbers. A routing path consists of a sequence of these AS numbers and since router tables contain paths to a large number of destinations we can construct a picture of the Internet at the autonomous system level by examining them. The process is similar to that for the traceroute method described in the previous section and depicted in Fig. 2.2. We must first obtain a set of router tables, which is normally done simply by asking router operators for access to their tables. Each router table contains a large number of paths starting from a single source (the router), and the union of the paths from many routers gives a good, though not complete, network snapshot in which the nodes are autonomous systems and the edges are the connections between autonomous systems. As with traceroute, it is important that the routers used be widely distributed across the network to avoid too much duplication of results, and the number of routers should be as large as possible to make the sampling of network edges as complete as possible. For example, the Routeviews Project,[5] a large BGP-based Internet mapping effort based at the University of Oregon, uses (again at the

[5] See `http://www.routeviews.org`

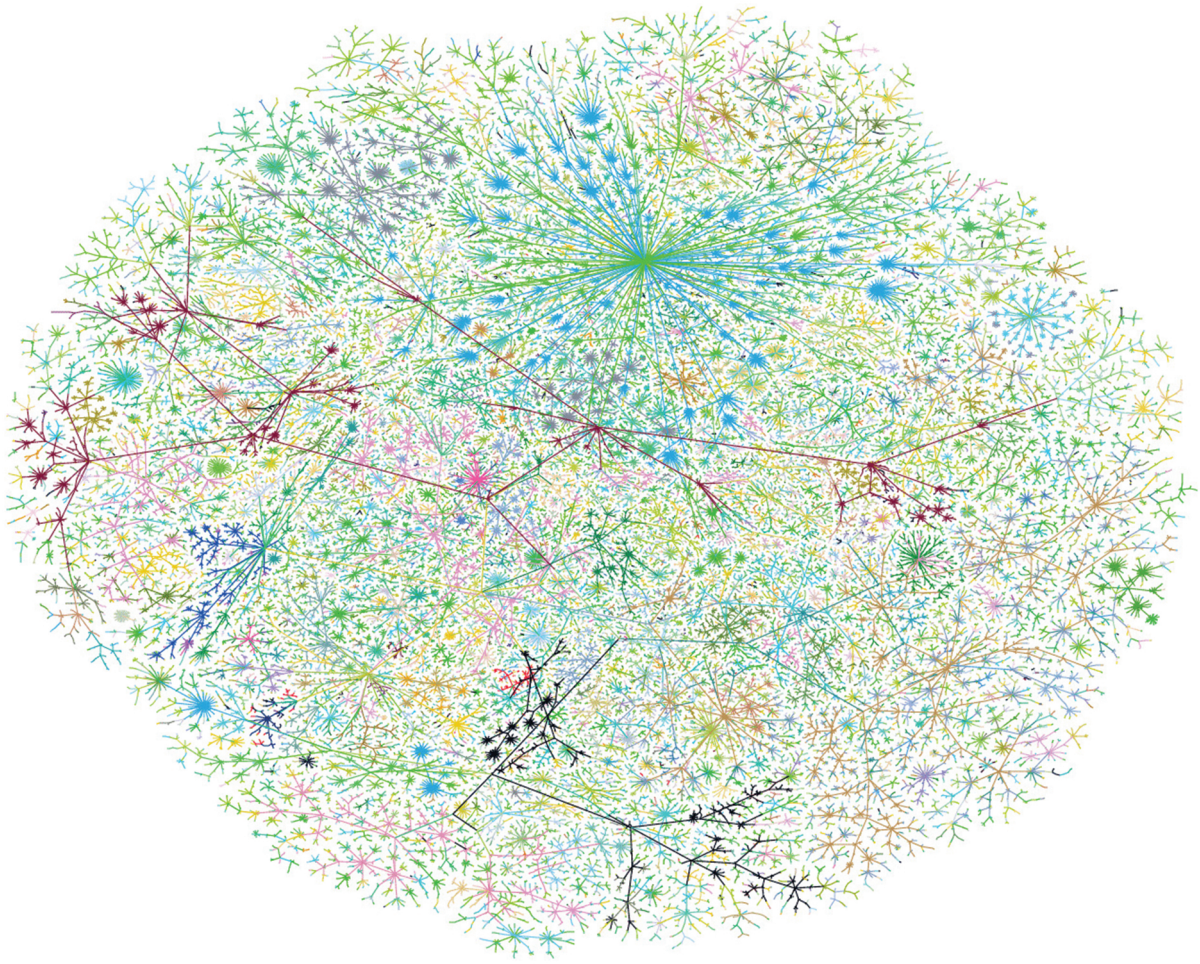


**Figure 2.3: The structure of the Internet at the level of autonomous systems.** The nodes in this network representation of the Internet are autonomous systems and the edges show the routes taken by data traveling between them. This figure is different from Fig. 1.1, which shows the network at the level of class C subnets. The picture was created by Hal Burch and Bill Cheswick. Patent(s) pending and Copyright Lumeta Corporation 2009. Reproduced with permission.

time of writing) a total of 501 source computers in 340 ASes around the world to measure the structure of the entire network every two hours.

Figure 2.3 shows a picture of the Internet at the AS level derived from routing tables. Qualitatively, the picture is similar to Fig. 1.1 for the class C subnet structure, but there are differences arising because class C subnets are smaller units than many autonomous systems and so Fig. 1.1 is effectively a

finer-grained representation than Fig. 2.3.

Using router-, subnet-, domain-, or AS-level structural data for the Internet, many intriguing features of the network's topology have been discovered in recent years [85, 102, 168, 323, 381, 384], some of which are discussed in later chapters of this book.

One further aspect of the Internet worth mentioning here is the geographic location of its nodes on the surface of the Earth. In many of the networks that we will study in this book, nodes do not exist at any particular position in real space—the nodes of a citation network, for instance, are not located on any particular continent or in any particular town. The nodes of the Internet, however, are by and large quite well localized in space. Your computer sits on your desk, a router sits in the basement of an office building, and so forth. Some nodes do move around, such as those representing mobile phones, but even these have a well-defined geographic location at any given moment. Things become a bit more blurry once the network is coarse-grained. The domain `umich.edu` covers large parts of the state of Michigan. The domain `aol.com` covers most of North America. These are somewhat special cases, however, being unusually large domains. The majority of domains have a well-defined location at least to within a few miles. Furthermore, tools now exist for determining, at least approximately, the geographic location of a given IP address, domain, or autonomous system. Examples include *NetAcuity*, *IP2Location*, *MaxMind*, and many others. Geographic locations are determined primarily by looking them up in one of several registries that record the official addresses of the registered owners of IP addresses, domains, or autonomous systems. These addresses need not in all cases match the actual location of the corresponding computer hardware. For instance, the domain `ibm.com` is registered in New York City, but IBM's principal operations are in California. Nonetheless, an approximate picture of the geographic distribution of the Internet can be derived by these methods, and there has been some interest in the results [477].

For a review of work on geographic networks of various kinds see Barthélemy [46].

Geographic placement of nodes is a feature the Internet shares with several other technological networks, as we will see in the following sections, but rarely with networks of other kinds.[6]

[6]Social networks are perhaps the main exception. In many cases people or groups have reasonably well-defined geographic locations and a number of studies have looked at how geography and network structure interact [285, 300, 374, 439].

## 2.2 The telephone network

The Internet is the best studied example of a technological network, at least as measured by the volume of recent academic work. This is partly because data on Internet structure are relatively easy to come by and partly because of intense interest among engineers, computer scientists, and the public at large. Other technological networks, however, are also of interest, including the telephone network and various distribution and transportation networks, and we look at some of these in the remainder of this chapter. Networks such as software call graphs and electronic circuits could also be considered technological networks and have been studied occasionally [174, 199, 334, 343, 485], but are beyond the scope of this book.

The telephone network—meaning the network of landlines and wireless links[7] that transmits telephone calls—is one of the oldest electronic communication networks still in use, but it has been studied relatively little by network scientists, primarily because of a lack of good data about its structure. The structure of the phone network is known in principle, but the data are largely proprietary to the telephone companies that operate the network and, while not precisely secret, they are not openly shared with the research community in the same way that Internet data are. We hope that this situation will change, although the issue may become moot in the not too distant future, as telephone companies are sending an increasing amount of voice traffic over the Internet rather than over dedicated telephone lines, and it may not be long before the two networks merge into one.

Some general principles of operation of the telephone network are clear however. By contrast with the Internet, the traditional telephone network is not a packet-switched network of the kind described in Section 2.1. Signals sent over the phone network are not disassembled and sent as sets of discrete packets the way Internet data are (though there are exceptions—see below). The telephone network is a *circuit-switched* network, which means that the telephone company has a number of lines or circuits available to carry telephone calls between different points and it assigns them to individual callers when those

[7] For most of its existence, the telephone network has connected together stationary telephones in fixed locations such as houses and offices using landlines. Starting in the 1980s, fixed telephones have been replaced by wireless phones ("mobile phones" or "cell phones"), but it is important to realize that even calls made on wireless phones are still primarily carried over traditional landline networks. The signal from a wireless phone makes the first step of its journey wirelessly to a nearby transmission tower, but from there it travels over ordinary phone lines. Thus, while the advent of wireless phones has had an extraordinary impact on society, it has had rather less impact on the nature of the telephone network.

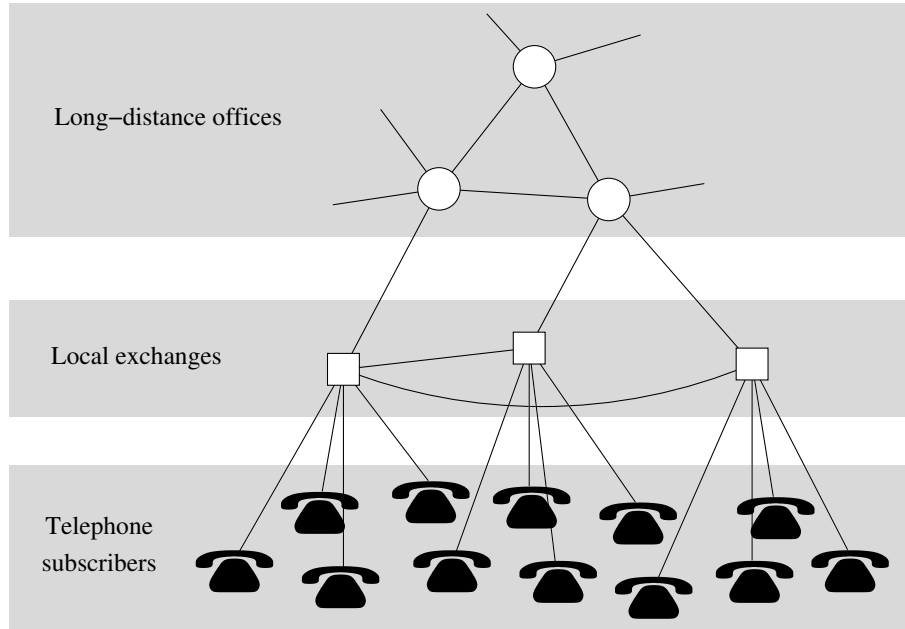



**Figure 2.4: A sketch of the three-tiered structure of a traditional telephone network.** Individual subscriber telephones are connected to local exchanges, which are connected in turn to long-distance offices. The long-distance offices are connected among themselves by trunk lines, and there may be some connections between local exchanges as well.

callers place phone calls. In the earliest days of telephone systems in the United States and Europe the "lines" actually were individual wires, one for each call the company could carry. Increasing the capacity of the network to carry more calls meant putting in more wires. Since the early part of the twentieth century, however, phone companies have employed techniques for *multiplexing* phone signals, i.e., sending many calls down the same wire simultaneously. The exception is the "last mile" of connection to the individual subscriber. The phone cable entering a house usually only carries one phone call at a time, although even that has changed in recent years as new technology has made it possible for households to have more than one telephone number and place more than one call at a time.

The basic form of the telephone network is relatively simple. Most countries with a mature landline telephone network use a three-tiered design, as shown in Fig. 2.4. Individual telephone subscribers are connected over local lines to

local telephone exchanges, which are then connected over shared "trunk" lines to long-distance offices, sometimes also called toll-switching offices. The long-distance offices are then connected among themselves by further trunk lines. The structure is, in many ways, rather similar to that of the Internet (Fig. 2.1), even though the underlying principles on which the two networks operate are different.

The three-level structure of the telephone network is designed to exploit the fact that most phone calls in most countries are local, meaning they connect subscribers in the same town or region. Phone calls between subscribers connected to the same local exchange can be handled by that exchange alone and do not need to make use of any trunk lines at all. Such calls are usually referred to as local calls, while calls that pass over trunk lines are referred to as trunk or long-distance calls. In many cases there may also be direct connections between nearby local exchanges that allow calls to be handled locally even when two subscribers are not technically attached to the same exchange.

The telephone network has had roughly this same topology for most of the past hundred years and still has it today, but many of the details about how the network works have changed. In particular, at the trunk level a lot of telephone networks are no longer circuit switched. Instead they are now digital packet-switched networks that work in a manner not dissimilar to the Internet, with voice calls being digitized, broken into packets, and transmitted over optical fiber links. Indeed, as mentioned, many calls are now transmitted digitally over the Internet itself, allowing phone companies to use the already existing Internet infrastructure rather than building their own. In many cases, only the "last mile" to the subscriber's telephone is still carried on an old-fashioned dedicated circuit, and even that is changing with the advent of digital and Internet telephone services and mobile phones. Nonetheless, in terms of geometry and topology the structure of the phone network is much the same as it has always been, being dictated in large part by the constraints of geography and the propensity for people to talk more often to others in their geographic vicinity than to those further away.

## 2.3 Power grids

The power grid is the network of high-voltage transmission lines that provide long-distance transport of electric power within and between countries. The nodes in a power grid correspond to generating stations and switching substations, and the edges correspond to the high-voltage lines. (Low-voltage local power delivery lines are normally not considered part of the grid, at least where network studies are concerned.) The topology of power grids is not difficult to

determine. The networks are usually overseen by a single authority and complete maps of grids are readily available. Very comprehensive data on power grids (as well as other energy-related networks such as oil and gas pipelines) are available from specialist publishers, either on paper or in electronic form, if one is willing to pay for them.

There is much of interest to be learned by looking at the structure of power grids [13, 20, 31, 125, 263, 378, 415, 466]. Like the Internet, power grids have a spatial element; the individual nodes each have a location somewhere on the globe, and their distribution in space is interesting from geographic, social, and economic points of view. Network statistics, both geographic and topological, may provide insight into the global constraints governing the shape and growth of grids. Power grids also display some unusual behaviors, such as cascading failures, which can give rise to surprising outcomes such as the observed power-law distribution in the sizes of power outages [140, 263].

However, while there is a temptation to apply network models of the kind described in this book to try to explain the behavior of power grids, it is wise to be cautious. Power grids are complicated systems. The flow of power is governed not only by geometry and simple physical laws, but also by detailed control of the phases and voltages across transmission lines, monitored and adjusted on rapid timescales by sophisticated computer systems and on slower timescales by human operators. There is evidence to suggest that network topology has only a relatively weak effect on power failures and other power-grid phenomena, and that good prediction and modeling of power systems requires more detailed information than can be gleaned from a network representation alone [234, 378].

## 2.4 Transportation networks

Another important class of technological networks are the transportation networks, such as airline routes and road and rail networks. The structure of these networks is not usually hard to determine. Airline networks can be reconstructed from published airline timetables, road and rail networks from maps. Geographic information systems (GIS) software can be useful for analyzing the geographic aspects of the data and there are also a variety of online resources providing useful information such as locations of airports.

One of the earliest examples of a study of a transportation network is the 1965 study by Pitts [387] of waterborne transport on Russian rivers in the Middle Ages. There was also a movement among geographers in the 1960s and 1970s to study road and rail networks, particularly focusing on the interplay between their physical structure and economics. The most prominent name

in the movement was that of Karel Kansky, and his book on transportation networks is a good point of entry into that body of literature [254].

More recently, a number of authors have produced studies applying new network analysis ideas to road, rail, air, and sea transportation networks [20, 34, 95, 198, 202, 224, 243, 290, 293, 324, 425, 426, 474]. In most of these studies the network nodes represent geographic locations and the edges represent routes. For instance, in studies of road networks the nodes usually represent road intersections and the edges roads. The study by Sen *et al.* [425] of the rail network of India provides an interesting counterexample. Sen *et al.* argue, plausibly, that in the context of rail travel what matters to most people is whether there is a direct train to their destination or, if there is not, how many trains they will have to take to get there. People do not care so much about how many stops there are along the way, so long as they don't have to change trains. Thus, Sen *et al.* argue, a useful network representation in the case of rail travel is one in which the nodes represent locations and two nodes are connected by an edge if a single train runs between them. Then the distance between two nodes in the network—the number of edges you need to traverse to get from A to B—is equal to the number of trains you would have to take. A better representation still (although Sen *et al.* did not consider it) would be a "bipartite network," a network containing two types of node, one representing the locations and the other representing train routes. Edges in the network would then join locations to the routes that run through them. The first, simpler representation of Sen *et al.* can be derived from the bipartite one by making a "projection" onto the locations only. Bipartite networks and their projections are discussed in Section 6.6.

## 2.5 Delivery and distribution networks

Falling somewhere between transportation networks and power grids are distribution networks, about which relatively little has been written within the field of networks research to date. Distribution networks include things like oil and gas pipelines, water and sewerage lines, and the routes used by the post office and package delivery companies. Figure 2.5 shows one example, the network of European gas pipelines, taken from a study by Carvalho *et al.* [96], who constructed the figure from data purchased from industry sources. In this network the edges are gas pipelines and the nodes are their intersections, including pumping, switching, and storage facilities and refineries.

If one is willing to interpret "distribution" in a loose sense, then one class of distribution networks that has been relatively well studied is river networks, though to be precise river networks are really collection networks rather than

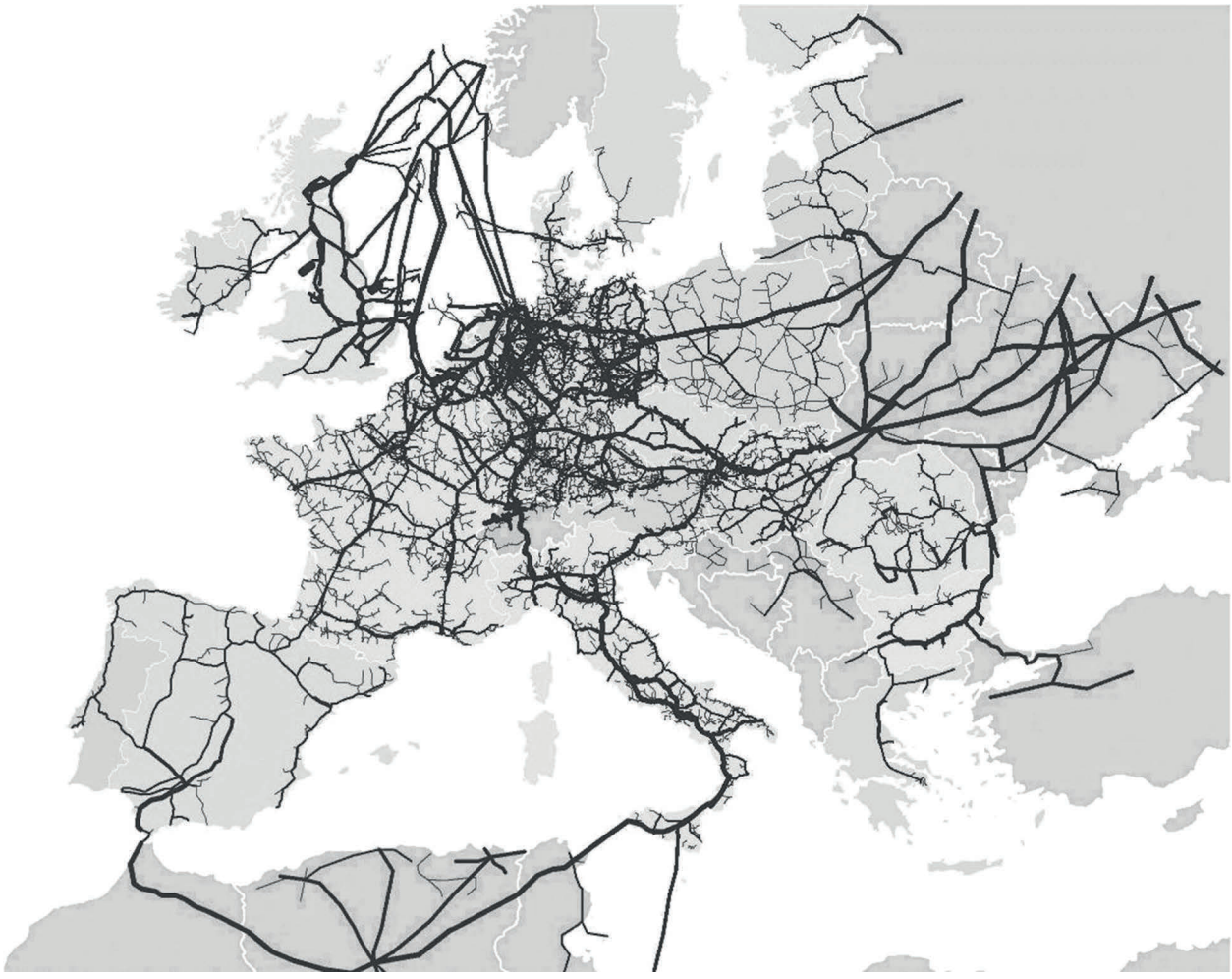


**Figure 2.5: The network of natural gas pipelines in Europe.** Thickness of lines indicates the sizes of the pipes. Reprinted with permission from R. Carvalho, L. Buzna, F. Bono, E. Gutierrez, W. Just, and D. Arrowsmith, Robustness of trans-European gas networks, *Phys. Rev. E* **80**, 016106 (2009). Copyright 2009 by the American Physical Society.

distribution networks. In a river network the edges are rivers or streams and the nodes are their intersections. As with road networks, no special techniques are necessary to gather data on the structure of river networks—the hard work of surveying the land has already been done for us by cartographers, and all we need do is copy the results from their maps. See Fig. 2.6 for an example.

The topological and geographic properties of river networks have been

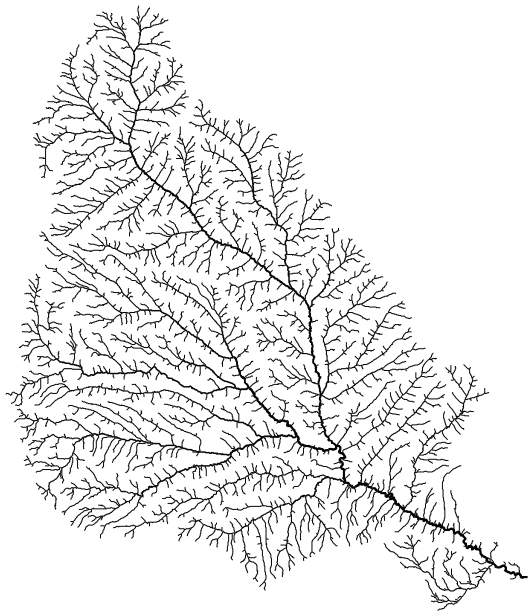


**Figure 2.6: Drainage basin of the Loess Plateau.** The network of rivers and streams on the Loess Plateau in the Shanxi province of China. The tree-like structure of the network is clearly visible—there are no loops in the network, so water at any point in the network drains off the plateau via a single path. Reproduced from Pelletier [386] by permission of the American Geophysical Union.

studied in some detail [143, 319, 407, 412]. Of particular note is the fact that river networks, to an excellent approximation, take the form of trees. That is, they contain no loops (if one disregards the occasional island midstream), a point that we discuss further in Section 6.8.

Similar in some respects to river networks are networks of blood vessels in animals, and their equivalents in plants, such as root networks. These too have been studied at some length. An early example of a mathematical result in this area is the formula for estimating the total geometric length of all edges in such a network by observing the number of times they intersect a regular array of straight lines [345]. This formula, whose derivation is related to the well-known "Buffon's needle" experiment for determining the value of $\pi$, is most often applied to root systems, but there is no reason it could not also be useful in the study of river networks or, with suitable modification, any other type of geographic network.

Also of note in this area is work on the scaling relationships between the structure of branching vascular networks in organisms and metabolic processes [39, 468, 469], an impressive example of the way in which an understanding of network structure can be parlayed into an understanding of the functioning of the systems the networks represent. We will see many more examples during the course of this book.

# Chapter 3

# Networks of information

*A discussion of information networks, with a particular focus on the World Wide Web and citation networks*

This chapter focuses on networks of information, networks consisting of items of data linked together in some way. Information networks are all, so far as we know, man-made, with perhaps the best known example being the World Wide Web, though many others exist and are worthy of study, particularly citation networks of various kinds.

In addition, there are some networks which could be considered information networks but which also have social-network aspects. Examples include networks of email communications, networks on social-networking websites such as Facebook or LinkedIn, and networks of weblogs and online journals. We delay discussion of these and similar examples to the following chapter on social networks, in Section 4.4, but they could easily have fitted in the present chapter also. The classification of networks as information networks, social networks, and so forth is a fuzzy one, and there are plenty of examples that, like these, straddle the boundaries.

## 3.1 The World Wide Web

Although by no means the first information network created, the World Wide Web is probably the example best known to most people and a good place to start our discussion in this chapter.

---

*Networks,* 2nd edition. Mark Newman, Oxford University Press (2018). 
DOI: 10.1093/oso/9780198805090.001.0001

As described in Chapter 1, the Web is a network in which the nodes are web pages, containing text, pictures, or other information, and the edges are the hyperlinks that allow us to navigate from page to page. The Web should not be confused with the Internet (Section 2.1), which is the physical network of data connections between computers; the Web is a network of links between pages of information.

Since hyperlinks run in one direction only, the Web is a directed network. We can picture the network with an arrow on each edge indicating which way it runs. Some pairs of web pages may be connected by hyperlinks running in both directions, which can be represented by two directed edges, one in each direction. Figure 3.1 shows a picture of a small portion of the web network, representing the connections between a set of web pages on a single website.

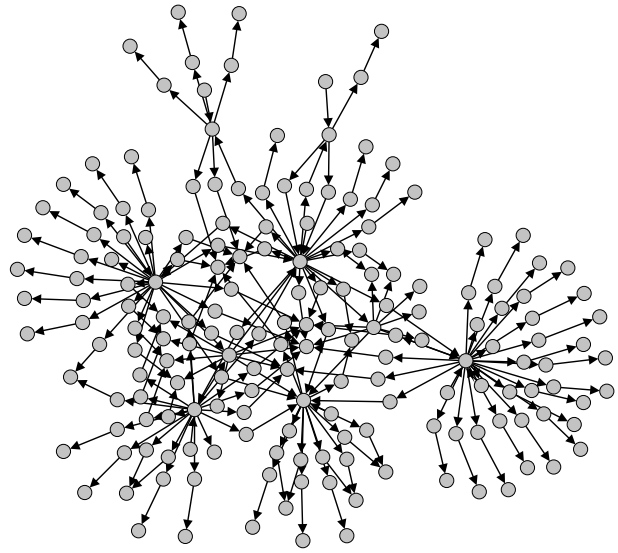


**Figure 3.1: A network of pages on a corporate website.** The nodes in this network represent pages on a website and the directed edges between them represent hyperlinks.

The World Wide Web was invented in the 1980s by scientists at the CERN high-energy physics laboratory in Geneva as a means of exchanging information among themselves and their co-workers, but it rapidly became clear that its potential was much greater [244]. At that time there were several similar information systems competing for dominance of the rapidly growing Internet, but the Web won the battle, largely because its inventors decided to give away for free the software technologies on which it was based—the Hypertext Markup Language (HTML) used to specify the appearance of pages and the Hypertext Transport Protocol (HTTP) used to transmit pages over the Internet. The Web's extraordinary rise is now a familiar part of history and most of us use its facilities at least occasionally, and in many cases daily. A crude estimate of the number of pages on the Web puts that number at around 50 billion at the time of the writing[1] and it is, almost certainly, the largest network that has been studied quantitatively by network scientists to date.

The structure of the Web can be measured using a *crawler*, a computer

[1] This is only the number of reachable static pages. The number of unreachable pages is difficult to estimate, and dynamic pages (see later) are essentially unlimited in number, although this may not be a very meaningful statement since these pages don't exist until someone asks for them.

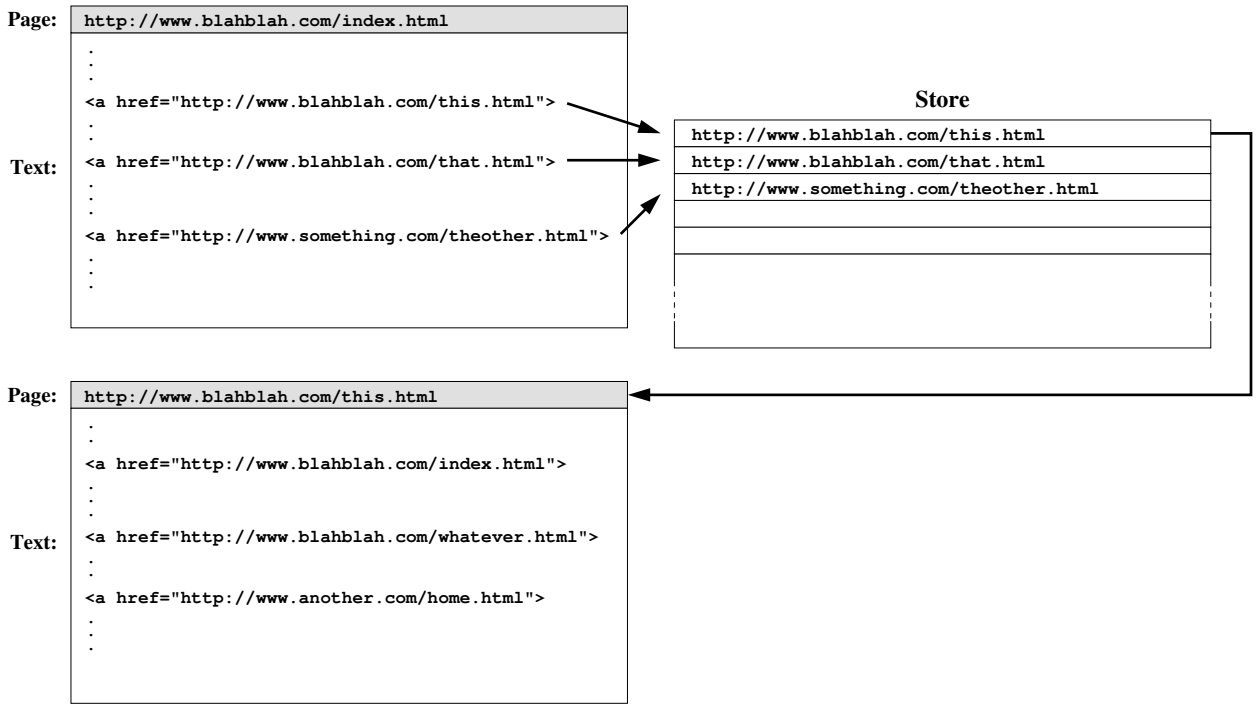



**Figure 3.2: The operation of a web crawler.** A web crawler iteratively downloads pages from the Web, starting from a given initial page. URLs are copied from the link tags in that initial page into a store. Once all links have been copied from the initial page, the crawler takes a URL from the store and downloads the corresponding page, then copies links from that, and so on.

Breadth-first search is discussed at length in Section 8.5.

program that automatically surfs the Web looking for pages. In its simplest form, the crawler performs a so-called breadth-first search on the web network, as shown schematically in Fig. 3.2. One starts from any initial web page, downloads the text of that page over the Internet, and finds all the links in the text. Functionally, a link consists of an identifying "tag"—a short piece of text marking the link as a link—and a *Uniform Resource Locator*, or URL, a standardized computer address that says how and where the linked web page can be found. By scanning for the tags and then copying the adjacent URLs a web crawler can rapidly extract URLs for all the links on a web page, storing them in memory or on a disk drive. When it is done with the current page, it takes one of the URLs from its store, uses it to locate a new page on the Web, and downloads the text of that page, and so the process repeats. If at any point the crawler encounters a URL it has seen before, then that URL is ignored and

not added to the store again, to avoid unnecessary duplication of effort. Only URLs that are different from those seen before are added to the store.

By repeating the process of downloading and URL extraction for a suitably long period of time, one can find a significant portion of the pages on the entire Web. No web crawler, however, finds all the pages on the Web, for a number of reasons. First, some websites forbid crawlers to examine their pages. Websites can place a file called `robots.txt` in their root directory that specifies which files, if any, crawlers can look at and may optionally specify that some crawlers are allowed to look at files while others are not. Compliance with the restrictions specified in a `robots.txt` file is voluntary, but in practice many crawlers do comply.

Second, many pages on the Web are dynamically generated: they are created on the fly by special software using, for instance, data from a database. Most large websites today, including many news, social media, retail, and corporate websites, as well as the web pages generated by search engines, fall into this category. Suppose, for instance, that you do a web search for "networks" using the Google search engine. Google does not keep a page of search results about networks (or anything else) just sitting on its computers, waiting for someone to ask for it. On the contrary, when you perform a search, the search engine rummages through its extensive database of web content (which it has found previously, using a web crawler) and makes a list of things that it believes will be useful to you. Then it creates a new web page containing that list and sends the page to your computer. The page of results you see when you search for something on Google is a dynamic page, generated automatically, and specifically for you, just a fraction of a second earlier.

As a result, the number of possible web pages that can be displayed as a result of a web search is so large as to be effectively infinite—as large as the number of different queries you could type into the search engine. When we are crawling the Web it is not practical for our crawler to visit all of these pages. The crawler must therefore make some choice about what it will look at and what it won't. One choice would be to restrict ourselves to static web pages—ones that are not generated on the fly. But it's not always simple to tell which pages are static, and besides, much useful information resides on the dynamic pages. In practice, the decisions made by crawlers about which pages to include tend to be fairly arbitrary, and it is not easy to guess which pages will be included in a crawl and which will not. But one can say with certainty that many will not and in this sense the crawl is always incomplete.

However, perhaps the most important reason why web crawls do not reach all the pages on the Web is that the network structure of the Web does not allow it. Since the Web is a directed network, not all pages are reachable from

a given starting point. In particular, it is clear that pages that have no incoming hyperlinks—pages that no one links to at all—can never be found by a crawler that follows links. Taking this idea one step further, it is also the case that a page will never be found if it is only linked to by pages that themselves have no incoming links. And so forth. In fact, the Web, and directed networks in general, have a special "component" structure, which we will examine in detail in Section 6.12.1, and most crawlers only find one part of that structure, the "giant out-component." In the case of the World Wide Web the giant out-component is estimated to occupy only about a half of all web pages and the other half of the Web is unreachable [84].[2]

Although we are interested in web crawlers as a tool for probing the structure of the Web so that we can study its network properties, this is not their main purpose. The primary use of web crawlers is to construct directories of web pages for search purposes. Web search engines such as Google indulge in web crawling on a massive scale to find web pages and construct indexes of the words and pictures they contain that can later be used to locate pages of interest to searchers. Because their primary interest is indexing, rather than reconstructing the network structure of the Web, search engine companies don't have any particular reason to take a good statistical sample of the Web and in network terms their crawls are probably quite biased. Still, many of them have graciously made their data available to academic researchers interested in web structure, and the data are good enough to give us a rough picture of what is going on. We will study a variety of features of the web network in subsequent chapters.

Web search, which itself raises some interesting network questions, is discussed in Section 18.1.

It isn't entirely necessary that we rely on search engine companies or other web enterprises for data on the structure of the Web. One can also perform one's own web crawls. There are a number of capable web crawler programs available for free on the Internet, including *wget*, *Nutch*, *GRUB*, and *Sphinx*. While most of us don't have the time or the facilities to crawl billions of web pages, these programs can be useful for crawling small sets of pages or single websites, and much useful insight and information can be acquired by doing so.

[2]Which web pages a crawler finds does depend on where the crawl starts. A crawler *can* find a web page with no incoming links, for instance, if (and only if) it starts at that page. In practice, however, the starting point has remarkably little effect on what a crawler finds, since most of what is found consists of the giant out-component mentioned above, whose content does not depend on the starting point.

## 3.2 Citation networks

A less well-known but much older information network is the network of citations between academic papers. Most papers reference one or more previous works, usually in a bibliography at the end of the paper, and one can construct a network in which the nodes are papers and there is a directed edge from paper A to paper B if A cites B in its bibliography. There are many reasons why one paper might cite another—to point out information that may be useful to the reader, to give credit for prior work, to highlight influences on the current work, or to disagree with the content of a paper. In general, however, if one paper cites another it is usually an indication that the contents of the earlier paper are relevant in some way to those of the later one, and hence citation networks are networks of relatedness of subject matter.

Quantitative studies of citation networks go back to the 1960s; perhaps the earliest is the 1965 study by Price [393]. Studies of citation networks fall within the field of information science, and more specifically within *bibliometrics*, the branch of information science that deals with the statistical study of publication patterns. The most common way to assemble a citation network is to do it by hand, simply typing the entries in the bibliographies of papers into a database from which a network can then be assembled. In the 1960s, when Price carried out his study, such databases were just starting to be created [200] and he made use of an early version of what would later become the Science Citation Index. Fifty years later, the Science Citation Index (along with its sister publications, the Social Science Citation Index and the Arts and Humanities Citation Index) is now one of the primary and most widely used sources of citation data. In recent years, it has moved from hand entry of bibliographic data to direct electronic submission of data by the journals, which makes for faster and more accurate database updates. Another database, Scopus, provides a competing but largely similar service. Both are professionally maintained and their coverage of the literature is reasonably complete and accurate, although the data are also quite expensive to purchase. Still, if one has the money, creating a citation network is only a matter of deciding which papers one wishes to include, using one of the databases to find the citations between those papers, and adding the appropriate directed edges to the network until it is complete.

Price's study is also the earliest we know of to find a power-law degree distribution in a network—see Section 10.4 for more discussion of this important phenomenon.

More recently, software systems for compiling citation indexes automatically without human oversight have started to appear. Perhaps the best known of these is *Google Scholar*, the academic literature arm of the Google search engine. Google Scholar crawls the Web to find manuscripts of papers in electronic form and then searches through those manuscripts to identify citations to other papers. This is a somewhat hit-or-miss operation because many papers are not

See Section 3.1 for a discussion of web crawlers.

on the Web or are not freely available, citations in papers have a wide variety of different formats and may include errors, and the same paper may exist in more than one place on the Web and possibly in more than one version. Nonetheless, enough progress has been made for Google Scholar to become a useful tool for the academic community. Other automated citation indexing projects include *Citebase*, which indexes physics papers, and *CiteseerX*, which indexes computer science.

As with web crawls, the original purpose of citation indexes was not to measure network structure. Citation indexes are assembled primarily to allow researchers to discover by whom a paper has been cited, and hence to find research related to a topic of interest. Nonetheless, data from citation indexes have been widely used to reconstruct the underlying networks and investigate their properties, and a number of large-scale studies of citation networks have appeared in recent years [101, 242, 294, 396–398, 404, 405].

Academic studies of the Web within the information sciences sometimes refer to hyperlinks as "citations," a nomenclature that emphasizes the close similarities between web and citation networks.

Citation networks are in many ways similar to the World Wide Web. The nodes of the network hold information in the form of text and pictures, just as web pages do, and the links from one paper to another play a role similar to hyperlinks between web pages, alerting the reader when information relevant to the topic of one paper can be found in another. Papers with many citations are often more influential and widely read than those with few, just as is the case with web pages, and one can "surf" the citation network by following a succession of citations from paper to paper just as computer users surf the Web.

Acyclic networks are discussed further in Section 6.4.1.

There is, however, at least one important difference between a citation network and the Web: a citation network is *acyclic*, while the Web is not. An acyclic network is one in which there are no closed loops of directed edges. On the World Wide Web, it is entirely possible to follow a succession of hyperlinks and end up back at the page you started at. On a citation network, by contrast, this is essentially impossible. The reason is that in order to cite a paper, that paper must already have been written. One cannot cite a paper that does not yet exist. Thus all the directed edges in a citation network point backward in time, from newer papers to older ones. If we follow a path of such edges from paper to paper, we will therefore find ourselves going backward in time, but there is no way to go forward again, so we cannot close the loop and return to where we started.[3]

See Fig. 6.3 for an illustration of a small acyclic network.

Citation networks have some interesting statistics. For instance, one study

[3] On rare occasions it occurs that an author will publish two papers simultaneously in the same volume of a journal and, with the help of the printers, arrange for each paper to cite the other, creating a cycle of length two in the network. Thus the citation network is not strictly acyclic, having a small number of short loops scattered about it.

found that about 47% of all papers have never been cited at all [404]. Of the remainder, 9% have one citation, 6% have two, and it goes down quickly after that. Only 21% of all papers have 10 or more citations, and just 1% have 100 or more. These figures are a consequence of the power-law degree distribution of the network—see Section 10.4.

The most highly cited paper of all, according to the Science Citation Index, is a 1951 paper by Lowry *et al.* [311], which has been cited more than 300 000 times.[4] Like most very highly cited papers, it is a methodological paper in molecular biology.

Citation networks of the type described so far are the simplest but not the only possible network representation of citation patterns. An alternative representation is the *cocitation network*. Two papers are said to be cocited if they are both cited by the same third paper. Cocitation is often taken as an indicator that papers deal with related topics and there is good evidence that this is a reasonable assumption in many cases. A cocitation network is a network in which the nodes represent papers and the edges represent cocitation of pairs of papers. By contrast with ordinary citation networks, the edges in a cocitation network are normally considered undirected, since cocitation is a symmetric relationship. One can also define a weighted cocitation network in which the edges have varying strengths: the strength of an edge between two papers is equal to the number of other papers that cite both.

Another related concept, although one that is less often used, is *bibliographic coupling*. Two papers are said to be bibliographically coupled if they cite the same other papers (rather than being cited by the same papers). Bibliographic coupling, like cocitation, can be taken as an indicator that papers deal with related material and one can define a strength or weight of coupling by the number of common citations between two papers. From the bibliographic coupling figures one can then assemble a bibliographic coupling network, either weighted or not, in which the nodes are papers and the undirected edges indicate bibliographic coupling.

### 3.2.1 PATENT AND LEGAL CITATIONS

The discussion of citation networks in the previous section focuses on citations between academic papers, but there are other types of citation also. Two of particular interest are citations between patents and between legal opinions.

Patents are temporary grants of ownership for inventions, which give their holders exclusive rights to control and profit from the protected inventions for a

[4] And it's been cited one more time now.

finite period of time. They are typically issued to inventors—either individuals or corporations—by national governments after a review process to determine whether the invention in question is original and has not been previously invented by someone else. In applying for a patent, an inventor must describe his or her invention in sufficient detail to make adequate review possible and present the case that the invention is worthy of patent protection. A part of this case typically involves detailing the relationship between the invention and other previously patented inventions, and in doing so the inventor will usually cite one or more previous patents. Citations may highlight dependencies between technologies, such as one invention relying for its operation on another, but more often patent citations are "defensive," meaning that the inventor cites the patent for a related previous technology and then presents an argument for why the new technology is sufficiently different from the old one to merit its own patent. Governments, in the process of examining patent applications, will routinely consider their similarity to previous inventions, and defensive citations are one way in which an inventor can fend off in advance possible objections that might be raised. Typically, there are a number of rounds of communication back and forth between the government patent examiner and the inventor before a patent application is finally accepted or rejected. During this process extra citations are often added to the application, either by the inventor or by the examiner, to document the further points discussed in their communications.

If and when a patent is finally granted, it is published, citations and all, so that the public may know which technologies have patent protection. Published patents thus provide a source of citation data that we can use to construct networks similar to the networks of citations between papers. In patent networks the nodes are patents, each identified by a unique patent number, and the directed edges between them are citations of one patent by another. Like academic citation networks, patent networks are acyclic, or nearly so, with edges running from more recent patents to older ones, although short loops can arise in the network in the not uncommon case that an inventor simultaneously patents a number of mutually dependent technologies.

The structure of patent networks reflects the organization of human technology in much the same way that the structure of academic citation networks reflects the organization of research knowledge. Patent citations have been less thoroughly studied than academic citations, but the number of studies has been growing in the past few years with the appearance of high-quality data sources, including US National Bureau of Economic Research database

of US patents[5] and the Google Patents search engine for worldwide patents.[6] There are a number of interesting technological and legal questions, for instance concerning originality of patented inventions, emerging technologies, and antitrust policy, that can be addressed by examining patent citation networks [106, 161, 216, 247].

Another class of citation networks that have begun to attract attention in recent years are legal citation networks. In countries where law cases can be decided by judges rather than juries, such as civil cases or appeals in Europe or the US, a judge will frequently issue an "opinion" after deciding a case, a narrative essay explaining his or her reasoning and conclusions. It is common practice in writing such an opinion to cite previous opinions issued in other cases in order to establish precedent, or occasionally to argue against it. Thus, like academic papers and patents, legal opinions form a citation network, with opinions being the nodes and citations being the directed edges, and again the network is approximately acyclic. The legal profession has long maintained indexes of citations between opinions for use by lawyers, judges, scholars, and others, and in recent years these indexes have made the jump to electronic form and are now available online. In the United States, for instance, two commercial services, LexisNexis and Westlaw,[7] provide detailed data on legal opinions and their citations. In the past few years a number of studies of the structure of legal citation networks have been published using data derived from these services [186, 187, 295, 314].

In principle it would be possible also to construct networks of cocitation or bibliographic coupling between either patents or legal opinions, but we are not aware of any studies yet published of such networks.

## 3.3 Other information networks

There are many other kinds of information networks, although none have attracted the same level of attention as the Web and citation networks. In the remainder of this chapter we briefly discuss a few examples of other networks.

[5] See `http://www.nber.org/patents`

[6] See `http://patents.google.com`

[7] Westlaw is owned and operated by Thomson Reuters, the same company that owns the Science Citation Index, while LexisNexis is owned by Elsevier, which also owns Scopus.

### 3.3.1 PEER-TO-PEER NETWORKS

*Peer-to-peer file-sharing networks* (sometimes abbreviated P2P) are a widely used form of computer network that combines aspects of information networks and technological networks. A peer-to-peer network is a network in which the nodes are computers containing information in the form, usually, of discrete files, and the edges between them are virtual links established for the purpose of sharing the contents of those files. The links exist only in software—they indicate only the intention of one computer to communicate with another should the need arise.

Peer-to-peer networks are typically used as a vehicle for distributed databases, particularly for the storage and distribution, often illegally, of music and movies, although there are substantial legal uses as well, such as local sharing of files on corporate networks or the distribution of software. (The network of router-to-router communications using the Border Gateway Protocol described in Section 2.1 is another less obvious example of a legitimate and useful peer-to-peer network.)

The point of a peer-to-peer network is to facilitate the direct transfer of data between computers belonging to two end users of the network, two "peers." This contrasts with the more common server–client model, such as that used by the World Wide Web, in which central server computers supply requested data to a large number of client machines. The peer-to-peer model is favored particularly for illicit sharing of copyrighted material because the owners of a centralized server can easily be obliged to deactivate the server by legal or law-enforcement action, but such actions are much more difficult when no central server exists. Eliminating central servers and the high-bandwidth connections they require also makes peer-to-peer networks economically attractive in applications such as software distribution.

On most peer-to-peer networks every computer is home to some information, but no computer has all the information in the network. If the user of a computer requires information stored on another computer, that information can be transmitted simply and directly over the Internet or over a local area network. This is a peer-to-peer transfer. No special infrastructure is necessary to accomplish it—standard Internet protocols are perfectly adequate to the task. Things get interesting, however, when one wants to *find* which other computer has the desired information. One way to do that is to have a central server containing an index of which information is on which computers. Such a system was employed by the early file-sharing network *Napster*, but a central server is, as we have said, susceptible to legal and other challenges, and such challenges

were in the end responsible for shutting Napster down.[8]

To avoid this problem, developers have turned to distributed schemes for searching and this is where network concepts come into play. In the simplest incarnation of the idea, computers form links to some number of their peers in such a way that all the computers together form a connected network. Again, a link here is purely a software construct—a computer's network neighbors in the peer-to-peer sense are merely those others with which it has agreed to communicate when the need arises.

When a user instructs his or her computer to search the network for a specific file, the computer sends out a message to its network neighbors asking whether they have that file. If they do, they arrange to transmit it back to the user. If they do not, they pass the message on to *their* neighbors, and so forth until the file is found. As we show in Section 18.2, where we discuss search strategies on peer-to-peer networks at some length, this algorithm works, but only on relatively small networks. Since it requires messages to be passed between many computers for each individual search, the algorithm does not scale well as the network becomes large, the volume of network traffic generated by searches eventually swamping the available data bandwidth. To get around this problem, modern peer-to-peer networks employ a two-tiered network topology of nodes and "supernodes," in which searches are performed only among the supernodes and ordinary nodes contact them directly to request searches be performed. More details are given in Section 18.2.

So what is the structure of a peer-to-peer network like? In many cases, unfortunately, not a lot is known since the software is proprietary and its owners are reluctant to share operational details. There have been a number of studies published of the early peer-to-peer network *Gnutella*, which was based on open-source software, meaning that the computer code for the software and the specification of the protocols it uses are freely available. By exploiting certain details of those protocols, particularly the ability for computers in the Gnutella network to "ping" one another (i.e., ask each other to identify themselves), researchers have been able to discover and analyze the structure of Gnutella networks [409, 442]. The networks appear to have approximately power-law degree distributions (see Section 10.4) and it has been suggested that this property could be exploited to improve search performance [6].

[8]The Napster name was later bought up by the music industry and is now the name of a legitimate online music service, although one that does not make use of peer-to-peer technology.

### 3.3.2 RECOMMENDER NETWORKS

*Recommender networks* represent people's preferences for things, such as for certain products sold by a retailer. Online merchants, for instance, may keep records of which customers bought which products and sometimes ask them whether they liked the products. Many large supermarket chains record the purchases made by their regular customers (usually identified by a small card with a barcode on it that is scanned when purchases are made) and so can work out which products each customer buys frequently.

We encountered bipartite networks previously in Section 2.4 and will study them further in Sections 4.5 and 6.6.

The fundamental representation of a recommender network is a "bipartite network," a network with two types of node, one representing the products or other items and the other representing the people, with edges connecting people to the items they buy or like. One can also add strengths or weights to the edges to indicate, for instance, how often a person has bought an item or how much he or she likes it, or the strengths could be made negative to indicate dislikes.

Recommender networks have been studied for many types of goods and products, including books, music, films, and others. Interest in recommender networks arises primarily from their use in *collaborative filtering systems*, also sometimes called *recommender systems*, which are computer algorithms that attempt to guess new items a person will like by comparing their past preferences with those of other people. If person A likes many of the same things as person B, for instance, and if person B likes some further item that A has never expressed an opinion about, then maybe (the theory goes) A would like that item too. A wide variety of computer algorithms have been developed for extracting conclusions of this type from recommender networks [406] and are used extensively by retailers to suggest possible purchases to their customers, in the hope of drumming up business. The website of the online retailer *Amazon.com*, for instance, has a feature that recommends items to customers based on their previously expressed preferences and purchases. And many supermarkets now print out personalized discount coupons at checkout for products that a customer has not bought in the past but might be interested to try.

Product recommendations of this kind are big business: the ability to accurately predict what customers will like can mean millions of dollars in extra sales for a large retailer, or the difference between a loyal customer and one who defects to a competitor. In 2006, the entertainment company *Netflix* offered a prize of one million US dollars for anyone who could create a recommender system able to predict viewers' opinions about movies and TV programs 10% more accurately than the company's existing system. A mere 10% may not seem like a big improvement, but for a business the size of Netflix, with millions of

users, it could translate into a substantial increase in profits, easily justifying the prize money. Moreover, it turned out to be no trivial task to beat the 10% threshold. It took almost three years before the prize was finally won in 2009 by a large collaborative team of US and European researchers.

Research on recommender networks has focused mainly on the development of better collaborative filtering algorithms, but it is reasonable to suppose that the success of these algorithms should depend to some extent on the structure of the recommender network itself, and there is therefore good reason to also study that structure. A few such studies have been published in the scientific literature [94, 220], but there is clearly room for further work.

### 3.3.3 Keyword indexes

Another type of information network, also bipartite in form, is the *keyword index*. An example is the index at the end of this book, which consists of a list of words or phrases, each accompanied by the numbers of the pages on which related information can be found. An index of this kind can be represented as a bipartite network, with two types of nodes representing words and pages, and an edge connecting each word to the pages on which it appears. In addition to their use in books, keyword indexes are routinely constructed as guides to other information collections, including sets of academic papers and the World Wide Web. The index constructed by a web search engine, as discussed in Section 3.1, is one example; it consists, at a minimum, of a set of words or phrases, with each word or phrase accompanied by a list of the web pages on which it occurs.

Indexes are of practical importance as a method for searching large bodies of information. Web search engines, for example, rely heavily on them to quickly find web pages that correspond to a particular query. However, indexes also have other, more sophisticated applications. They are used, for example, as a basis for techniques that attempt to find pages or documents that are similar to one another. Suppose one has a keyword index to a set of documents, consisting of a list of words and the documents they appear in. If we find that two documents contain a lot of the same keywords, it may be an indication that the two cover similar topics. A variety of computer algorithms for spotting such connections have been developed, typically making use of ideas very similar to those used in the recommender systems discussed in Section 3.3.2—the problem of finding documents with similar keywords is in many ways analogous to the problem of finding buyers who like similar products.

The identification of similar documents can be useful, for example, in constructing a search engine for searching through a body of knowledge. In a

standard index search, one typically looks up a keyword or set of keywords and gets a list of documents containing those words. Search engines that can tell when documents are similar may be able to respond more usefully because they can return documents that do not actually contain the keywords entered, but which are similar to documents that do. In cases where a single concept is called by more than one name, this may be an effective strategy for finding all the relevant documents.

In the context of document retrieval, the classic method for determining document similarity and performing generalized searches of this type is *latent semantic analysis*, which is based on the application of the matrix technique known as singular value decomposition to the bipartite network of keywords and documents [288]. A number of other competing methods have also been developed in recent years, using techniques such as non-negative matrix factorization [291, 292], latent Dirichlet allocation [63], and other probabilistic approaches [236].

As with recommender systems, it is reasonable to suppose that the success of methods for finding similar documents or improving searches using similarity information depends on the structure of the keyword index network, and hence that studies of that structure could generate useful insights. There has, however, been relatively little work on this problem so far within the network community, so there is plenty of room for future developments.

# Chapter 4

# Social networks

*A discussion of social networks and the empirical techniques used to probe their structure*

To most people the words "social network," if they mean anything, refer to online social media such as Facebook or Twitter. In the scientific study of networks, however, the phrase has a much broader meaning: a social network is any network in which the nodes represent people and the edges represent some form of connection between them, such as friendship. In this chapter we give a discussion of the origins and focus of the field of social network research and describe some of the types of networks studied and the techniques used to determine their structure. Sociologists have developed their own language for discussing social networks: they refer to the nodes, the people, as *actors* and the edges as *ties*. We will sometimes use these words when discussing social networks.

## 4.1 The empirical study of social networks

Interest in social networks goes back many decades. Indeed, among researchers studying networks sociologists have perhaps the longest and best established tradition of quantitative, empirical work. There are clear antecedents of social network analysis to be found in the literature as far back as the end of the nineteenth century, though the true foundation of the field is usually attributed to psychiatrist Jacob Moreno, a Romanian immigrant to America who in the 1930s became interested in the dynamics of social interactions within groups of

*Networks,* 2nd edition. Mark Newman, Oxford University Press (2018). 
DOI: 10.1093/oso/9780198805090.001.0001

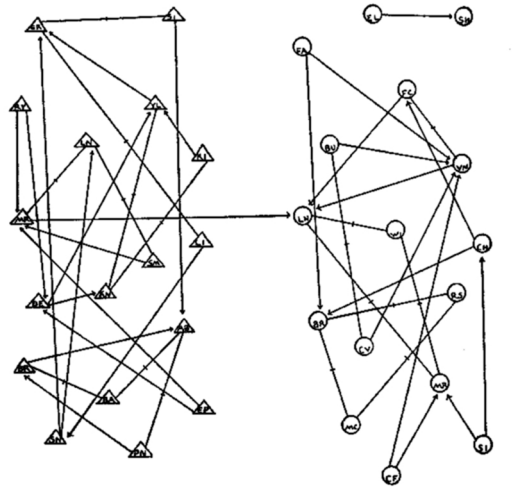


**Figure 4.1: Friendships between schoolchildren.** This early hand-drawn image of a social network, taken from the work of psychiatrist Jacob Moreno, depicts friendship patterns between the boys (triangles) and girls (circles) in a class of schoolchildren in the 1930s. Reproduced from [341] by kind permission of the American Society of Group Psychotherapy and Psychodrama.

people. At a medical conference in New York City in March 1933 he presented the results of a set of investigations he had performed that may have been the first true social network studies, and the work attracted enough attention to merit a column in the *New York Times* a few days later. The following year Moreno published a book entitled *Who Shall Survive?* [341] which, though not a rigorous work by modern standards, contained the seeds of the field of *sociometry*, which later became social network analysis.

The most startling feature of Moreno's work was a set of hand-drawn figures depicting patterns of interaction among various groups of people. He called these figures *sociograms* rather than social networks (a term not coined until about twenty years later), but in everything but name they are clearly what we now know as networks. Figure 4.1, for instance, shows a diagram from Moreno's book, depicting friendships among a group of schoolchildren. The triangles and circles represent boys and girls respectively, and the figure reveals, among other things, that there are many friendships among the boys and many among the girls, but only one between a boy and a girl. It is simple conclusions like this, that are both sociologically interesting and easy to see once one draws a picture, that rapidly persuaded social scientists that there was merit in Moreno's methods.

One of the most important things to appreciate about social networks is that there are many different possible definitions of an edge in such a network and the particular definition one uses will depend on what questions one is interested in answering. Edges might represent friendship between individuals, but they could also represent professional relationships, exchange of goods

or money, communication patterns, romantic or sexual relationships, or many other types of connection. If one is interested, for instance, in professional interactions between the boards of directors of major corporations, then a network of who looks at who else's Facebook page is probably not of much use. Moreover, the techniques one uses to probe different types of social interaction can be quite different, so that very different kinds of studies may be needed to address different kinds of questions. Direct questioning of experimental subjects is probably the most common method of determining the structure of social networks. We discuss it in detail in Section 4.2.

Another important technique, the use of archival records, is illustrated by a different early example of a social network study. In 1939 a group of ethnographers studying the effects of social class and stratification in the American south collected data on the attendance of social events by 18 women in a small town in Mississippi over a period of nine months [129]. Rather than relying on interviews or surveys, however, they assembled their data using guest lists from the events and reports in the society pages of the newspapers.[1] Their study, often referred to as the "Southern Women Study," has been widely discussed and analyzed in the networks literature in the decades since its first publication. The data can be represented as a network in which the nodes represent the women and two women are connected if they attended a common event. An alternative and more complete representation is as an "affiliation network" or "bipartite network," in which there are two types of node representing the women and the events, and edges connecting each woman to the events she attended. A visualization of the affiliation network for the Southern Women Study is shown in Fig. 4.2.

The use of archival and third-party records to reconstruct social networks is discussed in detail in Sections 4.4 and 4.5.

We encountered bipartite networks previously in Sections 2.4 and 3.3.2, and will study them in more detail in Sections 4.5 and 6.6.

One reason why this study has become so well known, in addition to its antiquity, is that the women were found by the researchers to split into two subgroups, tightly knit clusters of acquaintances with only rather loose between-cluster interaction. A classic problem in social network analysis is to devise a method or algorithm that can discover and extract such clustering from raw network data, and quite a number of researchers have made use of the Southern Women data as a test case for algorithm development.

Such is the power of social network analysis that its techniques have, since the time of Moreno and Davis *et al.*, been applied to an extraordinary variety of different communities, issues, and problems [79]: friendship and acquaintance patterns in local communities and in the population at large [54,55,261,333,447, 452] and among university students [446,479] and schoolchildren [169,338,400];

[1] They did also conduct some interviews, and made use of direct reports of attendance by observers. See Freeman [190] for a detailed discussion.

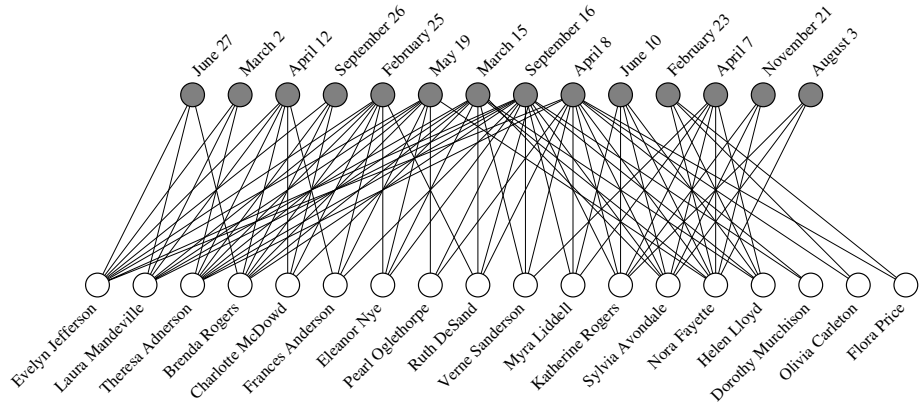



**Figure 4.2: The affiliation network of the "Southern Women Study."** This network (like all affiliation networks) has two types of node, the open circles at the bottom representing the 18 women who were the subjects of the study and the shaded circles at the top representing the social events they attended. The edges connect each woman to the events she attended. Data courtesy of L. Freeman and originally from Davis *et al.* [129].

contacts between business people and other professionals [117, 197]; boards of directors of companies [130, 131, 318]; collaborations of scientists [218, 219, 349], movie actors [20, 466], and musicians [206]; sexual contact networks [271, 305, 392, 411, 417] and dating patterns [52, 238]; covert and criminal networks such as networks of drug users [421] or terrorists [282]; historical networks [51, 377]; online communities such as Usenet [313, 431, 449] and Facebook [278, 298, 446, 452]; and social networks of animals [180, 315, 418, 419].

We will see examples of these and other networks throughout this book and we will give details as needed as we go along. The rest of the present chapter is devoted to a discussion of the different empirical methods used to measure social networks. The techniques described above, direct questioning of subjects and the use of archival records, are two of the most important, but there are several others that find regular use. This chapter does not aim to give a complete review of the subject—for that we refer the reader to specialized texts such as those of Wasserman and Faust [462] and Scott [424]—but the material here provides a good grounding for our further studies in the remainder of the book.

## 4.2 Interviews and questionnaires

The most common general method for gathering data on social networks is simply to ask people questions. If you are interested in friendship networks, you ask people who their friends are. If you are interested in business relationships you ask people who they do business with, and so forth. The asking may take the form of direct interviews with participants or the completion of questionnaires, either on paper or electronically. Indeed many modern studies, particularly telephone surveys, employ a combination of both interviews and questionnaires, wherein a professional interviewer reads questions from a questionnaire to a participant. By using a questionnaire, the designers of the study can guarantee that questions are asked, to a good approximation, in a consistent order and with consistent wording. By employing an interviewer to do the asking the study gains flexibility and reliability—interviewees often take studies more seriously when answering questions put to them by a human being—and interviewers may be given some latitude to probe interviewees when they are unclear, unresponsive, or confused. These are important considerations, since misunderstanding and inconsistent responses to survey questions are substantial sources of error [320]. By making questions as uniform as possible and giving respondents personal help in understanding them, these errors can be reduced. A good introduction to social survey design and implementation is given by Rea and Parker [403].

To measure social networks, surveys typically employ a *name generator*, a question or series of questions that invite respondents to name others with whom they have contact of a specific kind. For example, in their classic study of friendship networks among schoolchildren, Rapoport and Horvath [400] asked children to complete a questionnaire that included items worded as follows:

> My best friend at ____ Junior High School is:
> My second-best friend at ____ Junior High School is:
> My third-best friend at ____ Junior High School is:
> ⋮
> My eighth-best friend at ____ Junior High School is:

The blanks "____" in the questionnaire were filled in with the appropriate school name.[2] The list stopped at the eighth-best friend and many participants did not complete all eight.

Ideally all students within a school would be surveyed, although Rapoport

[2] A junior high school in the United States is a school for children aged approximately 12 to 14 years.

and Horvath reported that in their case a few were absent on the day the survey was conducted. Note that the survey specifically asks children to name only friends within the school. The resulting network will therefore record friendship ties within the school but none to individuals elsewhere. This is a common issue: it is highly likely that any group of individuals surveyed will have at least some ties outside the group and one must decide what to do with these ties. Sometimes they are recorded. Sometimes, as here, they are not. Such details can be important since statistics derived from survey results will often depend on the decisions made.

There are a number of points to note about the data produced by name generators. First, the network ties, friendships in the case above, are determined by one respondent nominating another. This is a fundamentally asymmetric process. Individual A identifies individual B as their friend. In many cases B will also identify A as *their* friend, but there is no guarantee that this will happen and it is not uncommon for nomination to go in only one direction. We normally think of friendship as a two-way relationship, but surveys suggest that this not always the case. As a result, data derived from name generators are often best represented as directed networks, networks in which edges run in a particular direction from one node to another. If two individuals nominate each other then we have two directed edges, one pointing in either direction. Each node in the network then has two degrees, an out-degree—the number of friends identified by the corresponding individual—and an in-degree—the number of others who identified the individual as a friend.

We encountered directed networks previously in Chapter 1, in our discussion of the World Wide Web, and they are discussed in more detail in Section 6.4.

Recall that the degree of a node is the number of connections it has—see Section 6.10 for a detailed discussion.

This brings us to another point about name generators. It is common, as in the example above, for the experimenter to place a limit on the number of names a respondent can give. In the study of Rapoport and Horvath this limit was eight. Studies that impose such a limit are called *fixed choice* studies. The alternative is a *free choice* study, which imposes no limit.

Limits are often imposed purely for practical purposes, to reduce the work of the experimenter. However, they may also help respondents understand what is required of them. In surveys of schoolchildren, for instance, there are some children who, when asked to name their friends, will patiently name all the other children in the entire school, even if there are hundreds of them. Such responses are not particularly helpful—almost certainly the children in question are employing a different definition of friendship from that employed by most of their peers and by the investigators.

However, limiting the number of responses is for most purposes undesirable. In particular, it clearly limits the out-degree of the nodes in the network, imposing an artificial and possibly unrealistic cut-off. As discussed in Chapter 1, an interesting property of many networks is the existence of hubs, rare

nodes of unusually high degree, which, despite being few in number, can sometimes have a dominant effect on the behavior of the network. By employing a name generator that artificially cuts off the degree, any information about the existence of such hubs is lost.

It is worth noting, however, that even in a fixed choice study there is normally no limit on the *in*-degree of nodes in the network; there is no limit to the number of times an individual can be nominated by others. And indeed in many networks it is found that a small number of individuals are nominated an unusually large number of times. Rapoport and Horvath [400] observed this in their friendship networks: while most children in a school are nominated as a friend of only a few others, a small number of popular children are nominated very many times. Rapoport and Horvath were some of the first scientists in any field to study quantitatively the degree distributions of networks, reporting and commenting extensively on the in-degrees in their friendship networks.

Not all surveys employing name generators produce directed networks. Sometimes we are interested in ties that are intrinsically symmetric between the two parties involved, in which case the edges in the network are properly represented as undirected. An example is networks of sexual contact, which are widely studied to help us understand the spread of sexually transmitted diseases [271,305,392,417]. In such networks a tie between individuals A and B means that A and B had sex. While participants in studies sometimes do not remember who they had sex with or may be unwilling to talk about it, it is at least in principle a straightforward yes-or-no question whether two people had sex, and the answer should not depend on which of the two you ask. In such networks therefore, ties are normally represented as undirected.

If individuals' responses differ too often, it is a sign that one's data are unreliable. Thus one may be able to estimate the level of measurement error in the data by comparing responses.

Surveys can and often do ask respondents not just to name those with whom they have ties but to describe the nature of those ties as well. For instance, questions may ask respondents to name people they both like and dislike, or to name those with whom they have certain types of contact, such as socializing together, working together, or asking for advice. For example, in a study of the social network of a group of medical doctors, Coleman *et al.* [117] asked respondents the following questions:

> Who among your colleagues do you turn to most often for advice?
> With whom do you most often discuss your cases in the course of an ordinary week?
> Who are the friends among your colleagues whom you see most often socially?

The names of a maximum of three doctors could be given in response to each question. A survey such as this, which asks about several types of interactions,

effectively generates data on several different networks at once—the network of advice, the discussion network, and so forth—but all built upon the same set of nodes. Networks such as this are sometimes called "multilayer" or "multiplex" networks.

Multilayer networks are discussed further in Section 6.7.

Surveys may also pose questions aimed at measuring the strength of ties, asking, for instance, how often people interact or for how long, and they may ask individuals to give a basic description of themselves: their age, income, education, and so forth. Some of the most interesting results of social network studies concern the extent to which people's choice of whom they associate with reflects their own background and that of their associates. For instance, you might choose to socialize primarily with others of a similar age to yourself, but turn for advice to those who are older than you.

The common tendency of people to associate with others who are similar to themselves in some way is called "homophily" or "assortative mixing," and we discuss it in detail in Sections 7.7 and 10.7.

The main disadvantages of network studies based on direct questioning of participants are that they are first laborious and second inaccurate. The administering of interviews or questionnaires and the collation of the responses is a demanding job that has been only somewhat eased by the use of computers and online survey tools. As a result, most studies have been limited to a few tens or at most hundreds of respondents—the 34-node social network of Fig. 1.2 is a typical example. It is a rare study that contains more than a thousand actors, and studies such as the National Longitudinal Study of Adolescent Health,[3] which compiled responses from over 90 000 participants, are very unusual and extraordinarily costly. Only a substantial public interest such as, in that case, the control of disease, can justify the expense of performing them.

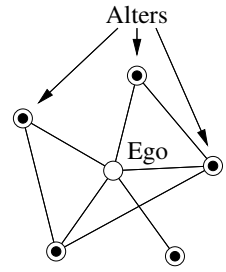
Data based on direct questioning may also be affected by biases of various kinds. Answers given by respondents are always to some extent subjective. If you ask people who their friends are, for instance, different people will interpret "friend" in different ways and thus give different kinds of answers, despite the best efforts of investigators to pose questions and record the answers in a uniform fashion. This problem is not unique to network studies. Virtually all social surveys suffer from such problems and a large body of expertise concerning techniques for dealing with them has been developed [320, 403]. Nonetheless, one should bear in mind when dealing with any social network derived from interviews or questionnaires the possibility of experimental bias in the data.

Experimental error in network measurements is discussed in detail in Chapter 9.

[3] See http://www.cpc.unc.edu/projects/addhealth

### 4.2.1 Ego-centered networks

Studies in which all or nearly all of the individuals in a community are surveyed, as described in the previous section, are called *sociometric* studies, a term coined by Jacob Moreno himself (see the discussion at the beginning of this chapter). Sociometric studies are the gold standard for determining network structure but, as discussed at the end of the preceding section, they are also very labor intensive and for large populations may be infeasible.




An ego-centered network consisting of an ego and five alters.

At the other end of the spectrum from sociometric studies lie studies of *personal networks* or *ego-centered networks*.[4] An ego-centered network is the network surrounding one particular individual, meaning that individual plus his or her immediate contacts. The individual in question is referred to as the *ego* and the contacts as *alters*.

Ego-centered networks are usually studied by direct questioning of participants, with interviews, questionnaires, or a combination of both being the instruments of choice (see Section 4.2). Typically, one constructs not just a single ego-centered network but several, centered on different egos drawn from the target population. In a telephone survey, for instance, one might call random telephone numbers in the target area and survey those who answer, asking them to identify others with whom they have a certain type of contact. Participants might also be asked to describe some characteristics both of themselves and of their alters, and perhaps to answer some other simple questions, such as which alters also have contact with one another.

Obviously, surveys of this type, and studies of ego-centered networks in general, cannot reveal the structure of an entire network. One receives snapshots of small local regions of the network, but in general those regions will not join together to form a complete social network. Sometimes, however, we are primarily interested in local network properties, and ego-centered network studies can give us good data about these. For example, if we wish to know about the degrees of nodes in a network—the numbers of ties people have—then a study in which a random sample of people are each asked to list their contacts may give us everything we need. (Studies probing node degrees are discussed more below.) If we also gather data on contacts between alters, we can estimate clustering coefficients (see Section 7.3). If we have data on characteristics of egos and alters we can measure assortative mixing (Sections 7.7 and 10.7).

An example of a study gathering ego-centered network data is the General

[4] Also called *egocentric* networks, although this term, which has its origins in social science and psychology, has taken on a different lay meaning which prompts us to avoid its use here.

Social Survey (GSS), a large-scale survey conducted every year in the United States starting in 1972 and every two years since 1994 [88]. The GSS is not primarily a social network study. The purpose of the study is to gather data about life in the United States, how it is changing, and how it differs from or relates to life in other societies. The GSS questionnaire contains a large number of parts, ranging from general questions probing the demographics and attitudes of the participants to specific questions about recent events, political topics, or quality of life. Among these many items, however, there are in each iteration of the survey a few questions about social networks. The precise number and wording of these questions changes from one year to another, but here are some examples from the survey of 1998, which was fairly typical:

> From time to time, most people discuss important matters with other people. Looking back over the last six months, who are the people with whom you discussed matters important to you? Do you feel equally close to all these people?
>
> Thinking now of close friends—not your husband or wife or partner or family members, but people you feel fairly close to—how many close friends would you say you have? How many of these close friends are people you work with now? How many of these close friends are your neighbors now?

By their nature these questions are of a "free choice" type, the number of friends or acquaintances the respondent can name being unlimited, although (and this is a criticism that has been leveled at the survey) they are also quite vague in their definition of friends and acquaintances, so people may give answers of widely varying kinds.

Another example of an ego-centered network study is the study by Bernard *et al.* [54, 55, 261, 326] of the degree of individuals in acquaintance networks (i.e., the number of people that people know). It is quite difficult to estimate how many people a person knows because most people cannot recall at will all those with whom they are acquainted and there is besides a lot of variation in people's subjective definition of "knowing." Bernard and co-workers came up with an elegant experimental technique for circumventing these difficulties. They asked study participants to read through a list of several hundred family names drawn at random from a telephone directory, and to count up how many people they knew with names appearing on the list. Each person with a listed name was counted separately, so that two acquaintances called Smith would count as two people. They were instructed to use the following precise definition of acquaintance:

Some care must be taken to match the selection of names to the community surveyed, since the frequency of occurrence of names shows considerable geographic and cultural variation.

> You know the person and they know you by sight or by name; you can contact them in person by telephone or by mail; and you have had contact with the person in the past two years.

(Of course, many other definitions are possible. By varying the definition, one could probe different social networks.) Bernard and co-workers then multiplied the counts reported by participants by a scaling factor to estimate the total number of acquaintances of each participant. For instance, if the random names used in the study accounted for 1% of the population, then one would multiply by 100 to estimate the total number of acquaintances.

Bernard and co-workers repeated their study with populations drawn from several different US cities and the results varied somewhat from city to city, but overall they found that the typical number of acquaintances, in the sense defined above, of the average person in the United States is about 2000. In the city of Jacksonville, Florida, for instance, they found a figure of 1700, while in Orange County, California they found a figure of 2025. Many people find these numbers surprisingly high upon first encountering them, perhaps precisely because we are poor at recalling all of the many people we know. But repeated studies have confirmed figures of the same order of magnitude, at least in the United States. In some other countries the figures are different. Bernard and co-workers repeated their study in Mexico City, for instance, and found that the average person there knows about 570 others.

## 4.3 Direct observation

An obvious method for constructing social networks is direct observation. Simply by watching interactions between individuals one can, over a period of time, form a picture of the networks of unseen ties that exist between those individuals. Most of us, for instance, will be at least somewhat aware of friendships or enmities that exist between our friends or co-workers. In direct observation studies, researchers attempt to develop similar insights about whatever population they are interested in.

Direct observation tends to be a labor-intensive method of study, so its use is usually restricted to small groups, and primarily ones with extensive face-to-face interactions in public settings. In Chapter 1 we saw one example, the "karate club" network of Zachary [479] (see Fig. 1.2 on page 5). Another is the study by Freeman *et al.* [193, 194] of the social interactions of a group of windsurfers, in which experimenters watched windsurfers on a beach in Orange County, California and recorded the length in minutes of every pairwise interaction among them. A large number of direct-observation network data

sets were compiled by Bernard and co-workers during the 1970s and 1980s as part of a lengthy study of the accuracy of individuals' perception of their own social situation [56, 58, 59, 259]. These included data sets on interactions among students, faculty, and staff in a university department, on members of a university fraternity,[5] on users of a teletype service for the deaf, and several other examples.

One arena in which direct observation is essentially the only viable experimental technique is studies of the social networks of animals, since clearly animals cannot be surveyed using interviews or questionnaires. Not all animal species form interesting social networks, but informative studies have been performed of, among others, monkeys [180, 418, 419], kangaroos [214], and dolphins [121, 315]. A common approach is to record instances of animal pairs engaging in recognizable social behaviors such as mutual grooming, courting, or close association, and then to declare ties to exist between the pairs that engage in these behaviors most often. Networks in which the ties represent aggressive behaviors have also been reported, such as networks of baboons [328], bison [310], deer [27], wolves [249, 455], and ants [116]. In cases where aggressive behaviors normally result in one animal's establishing dominance over another the resulting networks can be regarded as directed and are sometimes called *dominance hierarchies* [136, 137, 150].

## 4.4 Data from archival or third-party records

An increasingly important, voluminous, and often highly reliable source of social network data is archival records. Such records are, at least sometimes, relatively free from the vagaries of human memory and can be impressive in their scale, allowing us to construct networks of a size that would be unreachable by other methods. Archival records can also allow us to reconstruct networks that no longer exist, such as networks from the historical past.

A well-known, small-scale example of a study based on archival records is the work of Padgett and Ansell on the ruling families of Florence in the fifteenth century [377]. In this study the investigators looked at contemporaneous historical records to determine which among the Florentine families had trade relations, marriage ties, or other forms of social contact with one another. Figure 4.3 shows one of the resulting networks, a network of intermarriages between 15 of the families. It is notable that the Medici family occupies a central

[5] In American universities a "fraternity" is a combined social organization and boarding house for male students.

position in this network, having marriage ties with members of no fewer than six other families, and Padgett and Ansell conjectured that it was by shrewd manipulation of social ties such as these that the Medici rose to a position of dominance in Florentine society.

In recent years, researchers have used archival records to construct a wide variety of different networks, some of them very large. A number of authors, for example, have looked at email networks [156, 277, 450]. Drawing on email logs—automatic records kept by email servers of messages sent and received—it is possible to construct networks in which the nodes are people (or more correctly email addresses) and the directed edges between them are email messages. Exchange of email in such a network can be taken as a proxy for acquaintance between individuals, or we may be interested in patterns of email exchange for some other reason, such as understanding how information spreads through a community. Similar networks can also be constructed from patterns of text messaging or instant messaging using mobile phones [374, 439].

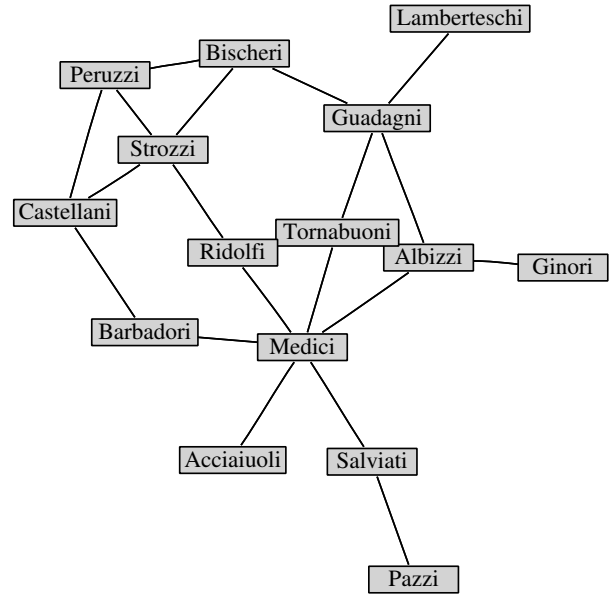



**Figure 4.3: Intermarriage network of the ruling families of Florence in the fifteen century.** In this network the nodes represent families and the edges represent ties of marriage between them. After Padgett and Ansell [377].

A network similar in some ways the email network is the *telephone call graph*, in which the nodes represent telephone numbers and directed edges between them represent telephone calls from one number to another. Call graphs can be constructed from call logs kept by telephone companies, and a number of studies have been performed in recent years, including some at the largest scales, with a million or more phone numbers [1, 10, 64, 233, 375, 401]. Studies of mobile phones have attracted particular attention because mobile phone data can reveal not only who calls whom but also potentially the geographic location of the phone users, providing a rare opportunity to construct networks with both detailed contact patterns and high spatial resolution [285, 374, 439]. Mobile phone data have also played a role in studies of face-to-face social interaction: if two phones are recorded as being in the same location at the same time one can perhaps conclude that their owners had face-to-face contact, and a number of studies have been conducted using assumptions of this kind [91, 154, 155, 439].

Telephone call graphs are quite distinct from the physical network of telephone cables discussed in Section 2.2 and the two should not be confused. Indeed, a call graph is to the physical telephone network roughly as an email network is to the Internet.

Email networks and telephone call graphs have another feature of partic-

ular interest: they are time-resolved—the date and time of each interaction is in principle known, allowing us to reconstruct after the fact the timing and duration of contacts between individuals if we have access to the appropriate data. Most of the sources of network data considered in this book are not time-resolved, but many networks do nonetheless change over time. Time-varying networks have been the focus of increasing research attention in recent years. We discuss them further in Section 6.7.

In sociology, studies of time-varying networks are sometimes called *longitudinal* studies, and you may occasionally encounter this term in the literature.

Recent years have also seen the rapid emergence of online social networking services, such as Facebook and LinkedIn, which, as a natural part of their operation, build records of connections between their participants and hence provide a rich source of archival network data. Some, such as Twitter, have made their data (or a part of it) publicly available, allowing researchers to study the corresponding networks [141, 208, 212]. Others are not publicly available but the companies involved have in some cases published analyses of their own networks, with some, such as Facebook, operating substantial internal research departments or inviting academic researchers to collaborate [28, 74, 278, 452]. Some online communities are not explicitly oriented towards networks or networking but can be studied using network techniques nonetheless. A number of researchers have looked, for instance, at networks of interactions between users of online dating sites [238, 297].

Weblogs, online diaries and journals, and other kinds of personal websites are another source of online social network data, although their popularity has waned somewhat in recent years. On these sites an individual or sometimes a group of people post their thoughts on topics of interest, often accompanied by links to other sites, and the sites and links form a directed network that lies, in terms of semantic content, somewhere between a social network and the World Wide Web: the links are often informational—the linker wishes to bring to his or her readers' attention the contents of the linked site—but there is a strong social element as well, since people often link to sites operated by their friends or acquaintances. The structure of the networks of links can be extracted using crawlers similar to those used to search the Web—see Section 3.1. Studies of weblogs and journals have been performed, for example, by Adamic and Glance [4] and MacKinnon and Warren [317].

## 4.5 Affiliation networks

An important special case of network data from archival records is the *affiliation network*. An affiliation network is a network in which actors are connected via their membership in groups of some kind. We saw one example at the start of this chapter, the Southern Women Study of Davis *et al.* [129], in which women

were connected via their common attendance at social events: the groups in that case were the attendees of the events. As we saw, the most complete representation of an affiliation network is a network with two types of nodes representing the actors and the groups, with edges connecting actors to the groups to which they belong—see Fig. 4.2 on page 50. In such a representation, called a "bipartite network" or "two-mode network," there are no edges connecting actors directly to other actors or groups to other groups, only actors to groups.

We study bipartite networks in more detail in Section 6.6.

Many examples of affiliation networks can be found in the literature. A famous case is the study by Galaskiewicz [197] of the CEOs of companies in Chicago in the 1970s and their social interaction via clubs that they attended: the CEOs are the actors and the clubs are the groups. Also in the business domain, a number of studies have been conducted of the networks formed by the boards of directors of companies [130,131,318], where the actors are company directors and the groups are the boards on which they sit. In addition to looking at the connections between directors in such networks, which arise as a result of their sitting on boards together, attention has also been focused on the connections between boards (and hence between companies) that arise as a result of their sharing a common director, a so-called board "interlock."

More recently, some extremely large affiliation networks have been studied in the mathematics and physics literature. Perhaps the best known example is the network of collaboration of film actors, in which the "actors" in the network sense are actors in the dramatic sense also, and the groups to which they belong are the casts of films. This network is the basis, among other things, for a well-known parlor game, sometimes called the "Six Degrees of Kevin Bacon," in which one attempts to connect pairs of actors via chains of intermediate costars, in a manner reminiscent of the small-world experiments of Stanley Milgram which we discuss in Section 4.6. The film actor network has, with the advent of the Internet, become very thoroughly documented and has attracted the attention of network analysts in recent years [20,40,466], although it is not clear whether there are any conclusions of real scientific interest to be drawn from its study.

Another example of a large affiliation network, one that holds more promise of providing useful results, is the coauthorship network of academics. In this network the actors are academic authors and the groups are the sets of authors of learned papers. Like the film actor network, this network is well documented, for instance via online bibliographic databases of published papers. Whether one is interested in papers published in journals or in more informal forums such as online preprint servers, excellent records now exist in most academic fields of authors and the papers they write, and a number of studies of the

corresponding affiliation networks have been published [43, 133, 196, 218, 219, 241, 347–349, 460].

## 4.6 The small-world experiment

A memorable and illuminating contribution to the social networks literature was made by the psychologist Stanley Milgram in the 1960s with his now-famous "small-world" experiments [333, 447]. Milgram was interested in quantifying the typical distance between actors in social networks. As discussed in Chapter 1, one can define the distance between two nodes in a network as the number of edges that must be traversed to go from one node to the other. There are mathematical arguments that suggest that this distance should be small for most pairs of nodes in most networks (see Section 11.7), a fact that was already well known in Milgram's time.[6] Milgram wanted to test this conjecture under real-world conditions and to do this he concocted the following experiment.

Milgram conducted several sets of small-world experiments. The one described here is the first and most famous, but there were others—see Refs. [275, 447].

Milgram sent a set of packages, 96 in all, to volunteer participants in the US town of Omaha, Nebraska, who were recruited via a newspaper advertisement. The packages contained an official-looking booklet, or "passport," emblazoned in gold letters with the name of Milgram's home institution, Harvard University, plus a set of written instructions. The instructions asked the participants to get the passport to a specified target individual, a friend of Milgram's who lived in Boston, Massachusetts, over a thousand miles away. The only information supplied about the target was his name (and hence indirectly the fact that he was male), his address, and his occupation as a stockbroker. But the passport holders were not allowed simply to send their passport to the given address. Instead they were asked to send it to someone they knew on a first-name basis, more specifically to the person in this category who they felt would stand the best chance of getting the passport to the intended target. Thus they might decide to send it to someone they knew who lived in Massachusetts, or maybe someone who worked in the financial industry. The choice was up to them. Whoever they did send the passport to was then to repeat the process, sending it to one of *their* acquaintances, and so forth, so that after a succession of steps the passport would, with luck, find its way into the hands of its intended recipient. Since every step of the process corresponded to the passport's changing hands between a pair of first-name acquaintances, the entire journey consti-

[6] Milgram was particularly influenced in his work by a mathematical paper by Pool and Kochen [389] that dealt with the small-world phenomenon and had circulated in preprint form in the social science community for some years when Milgram started thinking about the problem, although the paper was not officially published until some years later.

tuted a path along the edges of the social network formed by the set of all such acquaintanceships, and the length of the journey provided an upper bound on the distance through this network between the starting and ending individuals in the chain.

Of the 96 passports sent out, 18 found their way to the stockbroker target in Boston. (While this may at first sound like a low figure, it is actually quite high—recent attempts to repeat Milgram's work have resulted in response rates orders of magnitude lower [142].) Milgram asked participants to record in the passport each step of the path taken, so he knew how long each path was, and he found that the mean length of completed paths from Omaha to the target in Boston was just 5.9 steps. This result is the origin of the idea of the "six degrees of separation," the popular belief that there are only about six steps between any two people in the world.

The phrase "six degrees of separation" did not appear in Milgram's writing. It is more recent and comes from the title of a successful Broadway play by John Guare [221], later made into a film, in which the lead character discusses Milgram's work.

For a number of reasons this result is probably not very accurate. The initial recipients in the study were not chosen at random—they were volunteers who answered an advertisement—so they may not have been typical members of the population. At the very least, all of them were in a single town in a single country, which calls into question the extent to which the results of the study apply to the population of the world as a whole, or even to the population of the United States. Furthermore, Milgram used only a single target in Boston, and there is no guarantee this target was typical of the population either. Also we don't know that chains took the shortest possible route to the target. Probably they did not, at least in some cases, so the lengths of the chains provide only an upper bound on the actual distance between nodes. Moreover, most of the chains were never completed. Many passports were discarded or lost and never made their way to the target. It is reasonable to suppose that the chances of getting lost were greater for passports that took longer paths, and hence that the paths that were completed were a biased sample, having typical lengths shorter than the average.

For all of these reasons Milgram's results should be taken with a large pinch of salt. Even so, the fundamental conclusion that node pairs in social networks tend on average to be connected by short paths is now widely accepted. It has been confirmed directly in many cases, including for some very large social networks such as the entire network of Facebook friendships [452], and has moreover been shown to extend to many other (non-social) kinds of networks as well. Enough experiments have observed this "small-world effect" in enough networks that, whatever misgivings we may have about Milgram's particular technique, the general result is not seriously called into question.

Milgram's experiments also, as a bonus, revealed some other interesting features of acquaintance networks. For instance, Milgram found that most of

the passports that did find their way to the stockbroker target did so via just three of the target's friends. That is, a large fraction of the target's connections to the outside world seemed to be through only a few of his acquaintances, a phenomenon sometimes referred to as "funneling." Milgram called such well-connected acquaintances "sociometric superstars," and their existence has occasionally been noted in other networks also, such as collaboration networks [347], although not in some others [142].

A further interesting corollary of Milgram's experiment, never mentioned by Milgram himself, was highlighted many years later by Kleinberg [266, 267]: the fact that a moderate number of passports did find their way to the intended target person shows not only that short paths exist in the acquaintance network, but also that people are good at finding those paths. Upon reflection this is quite a surprising result. As Kleinberg has shown, it is possible and indeed common for a network to possess short paths between nodes but for them to be hard to find unless one has complete information about the structure of the entire network, which the participants in Milgram's studies did not. Kleinberg has conjectured that the network of acquaintances needs to have a special type of structure for the participants to find the paths they did with only limited knowledge of the network. We discuss his ideas in detail in Section 18.3.

Recently the small-world experiment has been repeated by Dodds *et al.* [142] using the modern medium of email. In this version of the experiment participants forwarded email messages to their acquaintances in an effort to get them to a specified target person about whom they were told a few basic facts. The experiment improved on Milgram's in terms of sheer volume, and also by having much more numerous and diverse target individuals and starting points for messages: 24 000 chains were started, most (though not all) with unique starting individuals, and with 18 different participating targets in 13 different countries. On the other hand, the experiment experienced enormously lower rates of participation than Milgram's, perhaps because the public is by now quite jaded in its attitude towards unsolicited mail. Of the 24 000 chains, only 384, or 1.5%, reached their intended targets, compared with 19% in Milgram's case. Still, the basic results were similar to those of Milgram. Completed chains had an average length of just over four steps. Because of their better data and more careful statistical analysis, Dodds *et al.* were also able to compensate for biases due to unfinished chains and estimated that the true average path length for the experiment was somewhere between five and seven steps—very similar to Milgram's result. However, Dodds *et al.* observed no equivalent of the "sociometric superstars" of Milgram's experiment, raising the question of whether their appearance in Milgram's case was a fluke of the particular target individual he chose rather than a generic property of social networks.

An interesting variant on the small-world experiment has been proposed by Killworth and Bernard [57,260], who were interested in how people "navigate" through social networks, and specifically how participants in the small-world experiments decided whom to forward messages to in the effort to reach the specified target. They conducted what they called "reverse small-world" experiments[7] in which they asked participants to *imagine* that they were taking part in a small-world experiment. A (fictitious) message was to be communicated to a target individual and participants were asked what they would like to know about the target in order to decide whom to forward the message to. The actual passing of the message never took place; the experimenters merely recorded what questions participants asked about the target. They found that three characteristics were sought overwhelmingly more often than any others, namely the name of the target, their geographic location, and their occupation—the same three pieces of information that Milgram provided in his original experiment. Some other characteristics came up with moderate frequency, particularly when the experiment was conducted in non-Western cultures or among minorities: in some cultures, for instance, parentage or religion were considered important identifying characteristics of the target.

The mechanisms of network search and message passing are discussed in greater detail in Section 18.3.

While the reverse small-world experiments do not directly tell us about the structure of social networks, they do give us information about how people perceive and deal with social networks.

## 4.7 Snowball sampling, contact tracing, and random walks

Finally in this chapter on social networks we take a look at a class of network-based techniques for sampling hidden populations.

Studies of some populations, such as drug users or illegal immigrants, present special problems to the investigator because the members of these populations do not usually want to be found and are often wary of giving interviews. Techniques have been developed, however, for sampling these populations by making use of the social networks that connect their members together. The most widely used such technique is *snowball sampling* [162, 188, 445].

Note that, unlike the other experimental techniques discussed in this chapter, snowball sampling is not intended as a technique for probing the structure of social networks. Rather, it is a technique for studying hidden populations

[7] Also sometimes called INDEX experiments, which is an abbreviation for "informant-defined experiment."

that relies on social networks for its operation. It is important to keep this distinction clear. To judge by the literature, some professional network scientists do not, a mistake that can result in erroneous conclusions and bad science.

Standard techniques such as telephone surveys often do not work well when sampling hidden populations. An investigator calling a random telephone number and asking if anyone on the other end of the line uses drugs is unlikely to receive a useful answer. The target population in such cases is small, so the chances of finding one of its members by random search are slim, and when you do find one they will very likely be unwilling to discuss the highly personal and possibly illicit topic of the survey with an investigator they have never met before and have no reason to trust.

So investigators probe the population instead by getting some of its members to provide contact details for others. The typical survey starts off rather like a standard ego-centered network study (see Section 4.2.1). You find one initial member of the population of interest and interview them about themselves. Then, upon gaining their confidence, you invite them also to name other members of the target population with whom they are acquainted. Then you go and find those acquaintances and interview them in turn, asking them also to name further contacts, and so forth through a succession of "waves" of sampling. Pretty soon the process "snowballs" and you have a large sample of your target population to work with.

Clearly this is a better way of finding a hidden population than random surveys, since each named individual is likely to be a member of the population, and you also have the advantage of an introduction to them from one of their acquaintances, which may make it more likely that they will talk to you. However, there are some serious problems with the method as well. In particular, snowball sampling gives highly biased samples. In the limit of a large number of waves, snowball sampling samples actors non-uniformly with probability proportional to their "eigenvector centrality" (see Section 7.1.2). In principle, knowing this, one could compensate for the non-uniformity by appropriately weighting the results, but in practice the limit of large number of waves is rarely reached, and in any case the eigenvector centrality cannot be calculated without knowledge of the complete contact network, which by definition we don't have, making correct weighting impossible. In short, snowball sampling gives biased samples of populations and there is little we can do about it. Nonetheless, the technique is sufficiently useful for finding populations that are otherwise hard to pin down that it has been widely used, biases and all, in studies over the past few decades.

Sometimes, in the case of small target populations, a few waves of snowball sampling may find essentially all members of a local population, in which case

the method can be regarded as returning data about the structure of the social network. If the contacts of each interviewed participant are recorded in the study, it should be possible to reconstruct the contact network when the study is complete. This has occasionally been done, although as noted above, the object is more often to exploit the social network to find the population than to study the network itself.

A technique closely related to snowball sampling is *contact tracing*, which is essentially a form of snowball sampling applied to disease incidence. Some diseases, such as tuberculosis and HIV, are considered in many countries to be sufficiently serious that, when someone is discovered to be carrying them, an effort must be made to track down all those who might also have been infected. Thus, when a patient tests positive for HIV, for instance, he or she will be questioned about recent sexual contacts, and possibly about other types of potentially disease-causing contacts, such as needle sharing if the patient is an injection drug user. Then health authorities will make an effort to track down the people so identified and test them for HIV also. The process is repeated with any who test positive, tracing their contacts as well, and so forth, until all leads have been exhausted. While the primary purpose of contact tracing is to curtail disease outbreaks and safeguard the health of the population, the process also produces data about the network through which a disease is spreading and such data have sometimes been used in scientific studies, particularly of sexually transmitted diseases, for which data may otherwise be hard to come by. Data from contact tracing studies display biases similar in type and magnitude to those seen in snowball sampling and should be treated with the same caution. Indeed, they may contain extra biases as well, since contacts are rarely pursued when an individual tests negative for the disease in question, so the sample is necessarily dominated by carriers of the disease, who are themselves usually a biased sample of the population at large.

There is another variant of snowball sampling that deals to some extent with the problems of bias in the sample. This is *random-walk sampling* [270, 445]. In this method one again starts with a single member of the target community and interviews them to determine their contacts. Then, however, instead of tracking down all of those contacts, one chooses one of them at random and interviews only that one at the next step. If the person in question cannot be found or declines to be interviewed, one chooses another contact, and the process is repeated. Initially it appears that this will be a more laborious process than standard snowball sampling, since one spends a lot of time determining the names of individuals one never interviews, but this is not the case. In either method one has to determine the contacts of each person interviewed, so the total amount of work for a sample of a given size is the same. It is, however, very