Chapter 1
Introduction and Overview

The systematic collection and analysis of data on networks of one form or another goes back at least to the 1930's in certain select areas of science, and in fact has subtle roots reaching back centuries further. However, during the decade surrounding the turn of the 21st century, network-centric analysis, as a general approach to scientific inquiry, has reached entirely new levels of prevalence and sophistication, with practitioners in fields now ranging from the physical and mathematical sciences to the social sciences and humanities. In this chapter we present a 'birds-eye' view of the area that is gradually coming to be known as 'network science,' starting with some background, continuing with a mosaic of examples, and finishing with a discussion of the organization and philosophy of this book.

1.1 Why Networks?

The oft-repeated statement that "we live in a connected world" perhaps best captures, in its simplicity, why networks have come to hold such interest in recent years. For example, from the 'small world' studies of Harvard sociologist Stanley Milgram [277] and later the play of Guare [188] comes the suggestion that we are each separated from any other person on the planet by at most six other people (i.e., 'six degrees'). And this concept arose even before the Internet and related inventions like email, chat-rooms, and blogs! Similarly, we see constantly in the popular press examples of the inter-connectedness of various human institutions (e.g., governments) and processes (e.g., economies), and also of humans and natural systems (e.g., in regards to the impact of humans on climate and the environment).

The image of a network – that is, essentially, something resembling a net – is a natural one to use to capture the notion of elements in a system and their interconnectedness. Note, however, that the term 'network' seems to be used in a variety of ways, at various levels of formality. The *Oxford English Dictionary,* for example, defines the word *network* in its most general form simply as "a collection of interconnected things." On the other hand, frequently 'network' is used inter-changeably

E.D. Kolaczyk, *Statistical Analysis of Network Data*, Springer Series in Statistics,
DOI 10.1007/978-0-387-88146-1_1, © Springer Science+Business Media, LLC 2009

with the term 'graph' since, for mathematical purposes, networks are most commonly represented in a formal manner using graphs of various kinds. In an effort to emphasize the distinction between the general concept and its mathematical formalization, in this book we will use the term 'network' in its most general sense above, and – at the risk of the impression of a slight redundancy – we will refer to a graph representing such a network as a 'network graph.' The discussion of the remainder of this chapter is at the level of networks, while technical material throughout the rest of the book is generally developed in reference to a specific network graph(s).

The seeds of network-based analysis in the sciences, particularly its mathematical foundation of graph theory, are often placed in the 1735 solution of Euler to the now famous Königsberg bridge problem, in which he proved that it was impossible to walk the seven bridges of that city in such a way as to traverse each only once. Since then, particularly since the mid-1800's onward, these seeds have grown in a number of key areas. For example, in mathematics the formal underpinnings were systematically laid, with König [235] cited as the first key architect. The theory of electrical circuits has always had a substantial network component, going back to work of Kirchoff, and similarly the study of molecular structure in chemistry, going back to Cayley. As the fields of operations research and computer science grew during the mid-1900's, networks were incorporated in a major fashion in problems involving transportation, allocation, and the like. And similarly during that time period, a small subset of sociologists, taking a particularly quantitative view towards the topic of social structure, began developing the use of networks in characterizing interactions within social groups.

Presently, examples of network-based analysis may now be found far beyond the traditional areas listed above, involving topics ranging from computer networking and the Internet to biology and gene networks to library science and webs of knowledge. Two important contributing factors to this growth are (i) an increasing tendency towards a systems-level perspective in the sciences, away from the reductionism that characterized much of the previous century, and (ii) an accompanying facility for high-throughput data collection, storage, and management. The quintessential example is perhaps that of the changes in biology over the past 10 to 20 years, during which the complete mapping of the human genome, a triumph of computational biology in and of itself, has now paved the way for fields like systems biology to be pursued aggressively, wherein a detailed understanding is sought of how the components of the human body, at the genetic level and higher, work together.

The focus of this book is on the statistical analysis of *network data.* More specifically, we aim to present a core set of methods and models for the analysis of measurements that are either of or from a system conceptualized as a network. Such data are collected daily in a host of different areas. Each area, naturally, has its own unique questions and problems under study. Nevertheless, from a statistical perspective, there is a methodological foundation emerging, composed of tasks and tools that are each common to some non-trivial subset of research areas involved with network science. It is our goal here to present this foundation.

Much of the challenge in analyzing network data stems from the fact that they involve, either explicitly or implicitly, quantities of a *relational* nature. As such, measurements are typically both high-dimensional and dependent. Additionally, such data are often substantial in quantity, and thus computational tractability is generally an issue not far from the surface when developing and using statistical methods and models in this area. The study of data that are either high-dimensional, dependent, or massive in quantity each is in itself currently an important topic of research in statistical theory and methods. In the analysis of network data, all three are often present in a unique fashion.

1.2 Examples of Networks

In order to better appreciate the nature of the statistical foundation emerging in the analysis of network data, it is useful to have some initial sense of the contexts in which networks arise, the scientific questions being asked, and the measurements being taken. While many such examples are to be found throughout the rest of the book, we present here in this section an initial glimpse, meant to provide somewhat of a mosaic picture. For convenience, and following Newman [296], the presentation is organized loosely into four classes of networks: technological, social, biological, and informational. These divisions are intended to be soft, and not hard, as many networks can be said to fall into more than one category.

1.2.1 Technological Networks

Arguably the networks most familiar to us are those of a technological nature (i.e., human constructions consciously created in a network form). Examples include communication networks (e.g., telephone networks or the Internet), transportation networks (e.g., networks of roads or rails, or networks of airline routes), and energy networks (e.g., networks for delivery of electricity or gas, or electrical circuits).

Some or all of the topology of such networks is often known by some entity. For example, telephone service providers maintain knowledge of the lines they lay and manufacturers of electrical circuits begin with blueprints of circuit designs. Connectivity may be in the form of a literal physical tie, such as a fiber optic cable or a gas line, or in the form of a virtual connection, such as a wireless link between a cellular phone and a nearby tower or the travel of an airplane between the airports of two cities. Interest in these networks is often focused on the flow of some corresponding 'commodity' across the network, be it Internet traffic packets, freight carried by trains, or units of electrical energy.

Consider the rather celebrated example of the Internet, which is essentially a network of digital devices communicating over wired and wireless connections via a set of communication protocols. Starting from comparatively modest beginnings, as

a 100-node research network in the mid-1970s, the Internet today is effectively a network of networks, linking on the order of hundreds of millions of devices and responsible for carrying the communication traffic that underlies everything from electronic banking transactions to email correspondence to music, video, and gaming entertainment.

Figure 1.1 shows a visual representation of a portion of the Internet, at a certain level of granularity – the sub-network known as Abilene. Abilene is part of the Internet2 project,[1] a research project devoted to development of the 'next generation' Internet. It serves as a so-called 'backbone' network for universities and research labs across the United States in a manner analogous to the federal highway system of roads. The 11 large-scale 'Core' nodes in this network correspond to regional network aggregation points, connected by systems of optical transportation technologies and routing devices, denoted here as links. In addition, connected to each Core node are additional nodes, such as 'Connector' nodes and 'Exchange Points.' The Connector nodes are network infrastructures through which local 'Participant' networks from universities and research labs access Abilene; the Exchange Points are similar, but instead serve to integrate Abilene with other 'Peer' Networks, such as similar networks in other countries. Note the clear hierarchical structure, which also continues down into the Participant networks themselves, and is replicated in the Peer networks, ultimately descending down to the laptops and such sitting in people's offices.

While most people largely take the Internet for granted, as a part of the infrastructure around which their daily lives are built, there is substantial interest in measuring and studying the Internet, in both the research and commercial communities. Network-oriented questions regarding the Internet tend to focus on those relating to its topology, the traffic it carries, the interaction of the two, and in turn the interaction of those with social and economic factors. For example, in regards to topology we may ask, "What does the Internet look like?" "How big is it?" and "What are its structural characteristics?" In terms of traffic, questions include "How much traffic is flowing across the network?" "How can I distinguish between 'normal' and 'anomalous' traffic?" and "Does my network have the capacity to meet anticipated demands?" See the book by Crovella and Krishnamurthy [105], for example.

In order to answer questions like these, measurements are taken in the Internet in a variety of active and passive manners. For example, it is possible to actively probe the Internet with small packets of traffic and register the responses that return to the sender as a result of communication protocols, which provides information on the routes the packets traveled and hence some insight into Internet topology. Conversely, it is also possible to 'sit' passively on an Internet link and monitor the traffic flowing by to a greater or lesser extent, depending on the granularity of information desired. We will see statistical topics of particular relevance to measurements like these in Chapter 3 (i.e., mapping networks), Chapter 5 (i.e., network sampling bias), Chapter 7 (i.e., Internet topology identification), Chapter 8 (i.e., spread of epidemics in a network), and Chapter 9 (i.e., analysis of network flow data).

[1] *http://www.internet2.edu/*

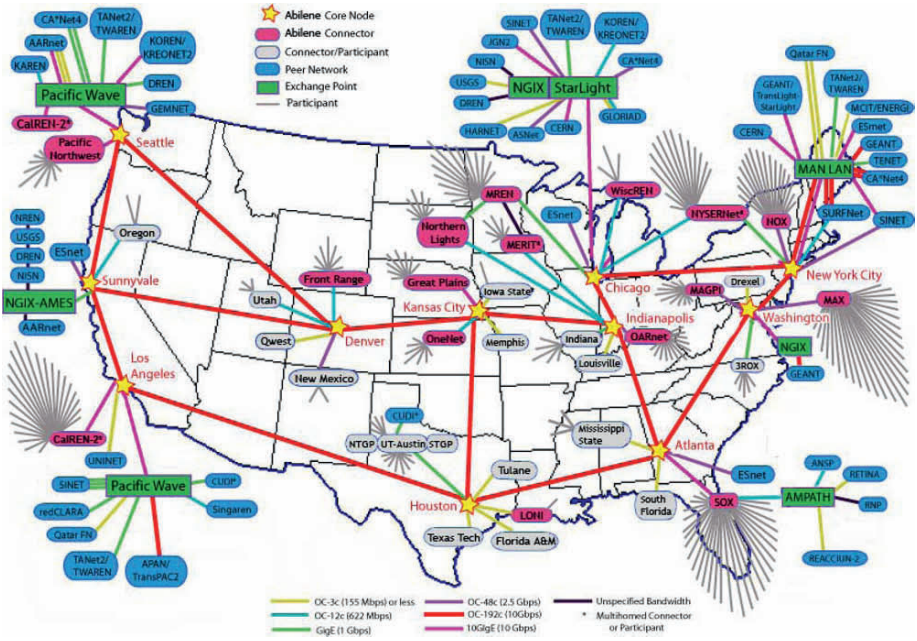

Fig. 1.1 Depiction of the Abilene network in the Internet. Different nodes represent various forms of network 'entities', while different colors of links indicate various levels of communication bandwidth. Note that some node names appear more than once, corresponding to the phenomena of 'multi-homing', wherein a given network connects to another at more than one location. Figure courtesy of Sucharita Gopal.

1.2.2 Social Networks

A class of networks with one of the longest histories of systematic study, dating back to at least the 1930's, is that of social networks (i.e., networks representing the interactions among a collection of social entities or 'actors'). Such entities typically are people or groups of people, but sometimes are non-human, such as animals. The type of interactions considered in this area varies and is constrained in part by the unit and nature of the social entities involved. Examples of social interactions include friendships among people, membership of people in larger social groups (e.g., clubs, companies, etc.), contacts between people (e.g., sexual contacts, meetings between members of terrorist cells, etc.), cooperation on a common endeavor, and the exchange of resources. Specific examples of social networks include networks of friendships among school children, sexual contacts within a community, corporate alliances among businesses, email exchanges between individuals, co-authorship on scientific articles, and trade agreements among nations.

The study of such networks is of particular interest to, and has traditionally been the province of, researchers in social sciences like sociology, anthropology, and psychology, although this interest is increasingly shared now by researchers in a number

of other areas, such as business and public health. The focus in these areas typically is on social structure and the quantitative characterization and analysis of such structure, which they aim to accomplish through the measurement of social interactions and the analysis of the resulting social network topologies. Questions of interest include "Who interacts with whom and what factors influence the tendency to interact?" "Which interactions are mutual?" "Are friends of friends also friends?" "What social groups, if any, exist in the network?", "Who are the power brokers?" "Who is central to the network and who is peripheral?" and "Which actors are similar in the roles they play?"

Figure 1.2 shows a visual representation of a particular social network, the so-called 'karate club network' of Zachary [411]. Nodes represent members of a karate club observed by Zachary for roughly two years during the 1970's, and links connecting two nodes indicate social interactions between the two members (thickness of the links reflects relative frequency of interactions). This dataset is somewhat unique in that Zachary had the curious fortune (from a scientific perspective) to witness the club split into two different clubs during his period of observation, due to a dispute between the head teacher and an administrator. It was originally published in conjunction with a model for information flow in small groups in the presence of conflict and fission. It has since become a favorite among researchers developing algorithms for detection of social subgroups, since the truth of membership in the two subgroups is a known quantity.

Social network scientists often face unique measurement challenges. Potential difficulties include the identification of social entities of interest, their willingness to be recruited into studies, and possible sources of bias in their response. For example, drug addicts or sex workers, two subpopulations of considerable interest to those studying the impact of social structure on the spread of the AIDS virus, are not necessarily obvious to identify in the general population. Once identified, they may have serious qualms about participating in a study. And once having agreed, in principle, to participate, they may be understandably reluctant to fully disclose information on, say, the sharing of hypodermic needles or their sexual partners. All of these issues, and others like them, can have an important impact on the statistical analysis of such data.

In recent years, the Internet has begun to have a fascinating impact on the field of social network analysis, due both to the potential for large-scale data acquisition and storage and the actual types of social interactions facilitated by the Internet. Examples of networks whose study is impacted in this manner include networks of email exchanges or phone calls, which can be measured by exploiting the existing computer infrastructure in the underlying technological networks, or networks of scientific collaborations, now relatively easily compiled on large scales because of the prevalence of electronic publication and archiving. Many of these networks are significantly larger than the size of typical social networks studied in the past. However, this characteristic can lead to its own issues. For example, the sheer magnitude of emails sent or calls made daily on a service provider's network quickly leads to concerns about data volume, which can make even seemingly simple tasks of network summary and visualization highly nontrivial.

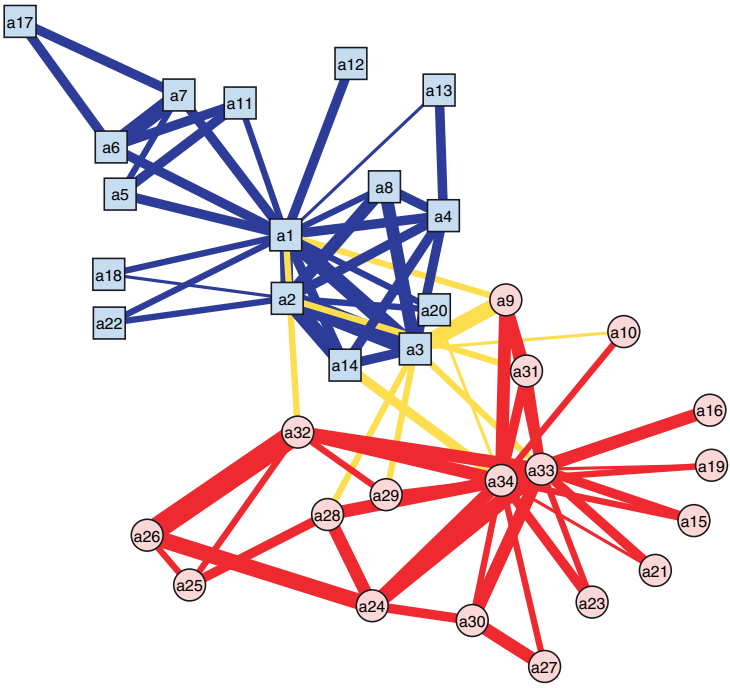

Fig. 1.2 Zachary's 'karate club' network. Subgroups, centered around actors 1 and 34, are indicated by the coloring and shape of their nodes, using blue squares and red circles, respectively. Links between actors within the same subgroup are colored similar to their nodes, while links between actors of different subgroups are shown in yellow.

We will encounter various social network tools and other topics of relevance to social network analysis in Chapter 3 (i.e., mapping networks), Chapter 4 (i.e., descriptive analysis of observed network structure), Chapter 5 (i.e., the sampling of network graphs), and Chapter 6 (i.e., the modeling of network graphs).

1.2.3 Biological Networks

Networks are a natural and commonly used tool for representing the internal workings of biological systems, at all different scales. For example, intra-cellular networks of interest include those describing the regulatory behavior among genes, the physical affinity for binding among proteins, the participation of metabolites together in biochemical processes, and combinations thereof. Similarly, a well-known example of an inter-cellular network is a network of neurons. On the other hand, networks describing interactions among complete organisms include ecological net-

works, such as those describing predator-prey relationships, and epidemiological networks, characterizing the spread of disease in a population.

As an illustration of a biological network, consider the network representation in Figure 1.3, which characterizes the cellular-level biomolecular process underlying the circadian clock mechanism in the organism *Drosophila melanogaster* (fruit fly). The 'circadian rhythm,' a 24-hour periodic cycle common to most living beings, is shown here as being driven by a feedback loop that creates a corresponding accumulation and decay in the quantities of two proteins, *Per* and *Tim.* Both the proteins and the genes that code for the proteins are indicated in the diagram, as rectangular and circular nodes, respectively, while the links indicate various steps in the overall process, including DNA translation and transcription, phosphorylation, suppression, and even physical movement across the nuclear membrane (drawn as a vertical dashed line).

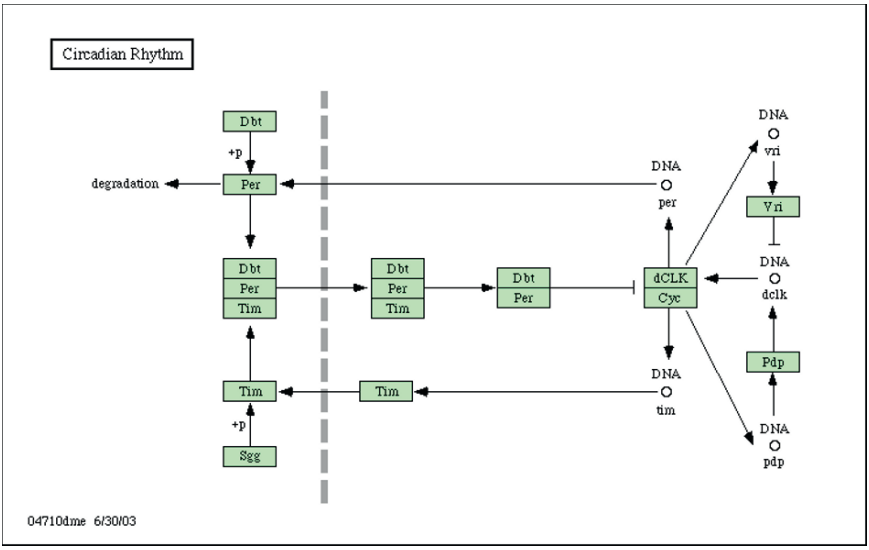

Fig. 1.3 Network representation of the circadian clock mechanism in *Drosophila melanogaster* (fruit fly), as of June 30, 2003, from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2].

Historically, the construction of such networks has been tremendously time-consuming. However, since roughly the mid-1990's, with the advent of high-throughput measurement techniques in genomics, the entire nature of this task has changed dramatically. The flood of data brings with it the potential to infer relational information like that in Figure 1.3 on a large scale in a semi- to fully-automated fashion. As a result, substantial energy is being focused on creating system-wide network representations of interactions among the different relevant biological elements (e.g., genes, proteins, etc.), and on the use of such networks for discovering higher-level phenomena associated with cellular processes. Examples of such tasks

range from the basic 'mapping' of the underlying biological system, to the search for meaningful clusters or sub-networks of system elements (i.e., 'motifs'), to the inference of roles in cellular function (e.g., protein function prediction) and the prediction of the behavior of the overall system in the face of external influences (e.g., the response to new cancer drug treatments).

Not surprisingly, the nature of the data collected on biological networks and the manner in which they are analyzed and used vary widely with the nature of the underlying biological system being studied and our ability to obtain relevant measurements. We will encounter a variety of examples of biological networks, at various scales, and topics relevant to their study in Chapter 3 (i.e., mapping networks), Chapter 5, (i.e., network sampling), Chapter 6 (i.e., detection of network motifs), Chapter 7 (i.e., inference of networks), and Chapter 8 (i.e., protein function prediction and modeling of epidemiological processes).

1.2.4 Information Networks

Of particular use in this modern 'information age,' although by no means new, are information networks (i.e., networks describing relationships among elements of information). Standard examples include networks of citations between academic journals or papers, networks of co-authorship on papers, or networks indicating semantic relationships (e.g., synonym, antonym, etc.) between words or concepts. In addition, the Internet has helped spawn a number of well-known classes of information networks. The pre-eminent example is the World Wide Web (WWW), in which nodes typically are web pages and edges indicate the referencing of one page by another. Another class of Internet-related information networks are peer-to-peer (i.e., 'P2P'), networks, in which nodes are typically Internet users and links indicate the exchange of content (e.g., music or movies) through an associated network protocol (e.g., Napster, Gnutella, KaZaa, etc.).

The 'mapping' of information networks in an informative fashion is usually a non-trivial task of significant interest in itself, particularly given their often massive size. Additionally, there is generally strong interest in questions regarding the structure of such networks, including which nodes are linked to many other nodes (e.g., "Who are the most highly cited authors within the mathematical sciences literature?"), whether certain tightly inter-woven subgraphs may be found (e.g., "How does the content of web pages induce clustering on the WWW?"), and the manner in which network size and structure change over time (e.g., "What are the dynamics of the lifetime of a scientific innovation?").

As an illustration of an information network, consider the network depicted in Figure 1.4, which is an example of an important class of sub-networks of the WWW called 'web-logs' or simply 'blogs'. Blogs are a form of Web authorship, primarily textual in nature but increasingly more multimedia based. The corresponding web page(s) of a blog consists of a set of entries, often on a particular topic(s), that are archived in reverse chronological order and usually updated frequently. 'Blogging,'

as the act of maintaining a blog is called, has come to refer to a range of activities that include personal journal keeping, citizen reporting, information dissemination, and social interaction (e.g., support groups, political action, etc.). In fact, blog networks can be viewed not only as information networks, but often also as social networks representing so-called 'virtual communities.'

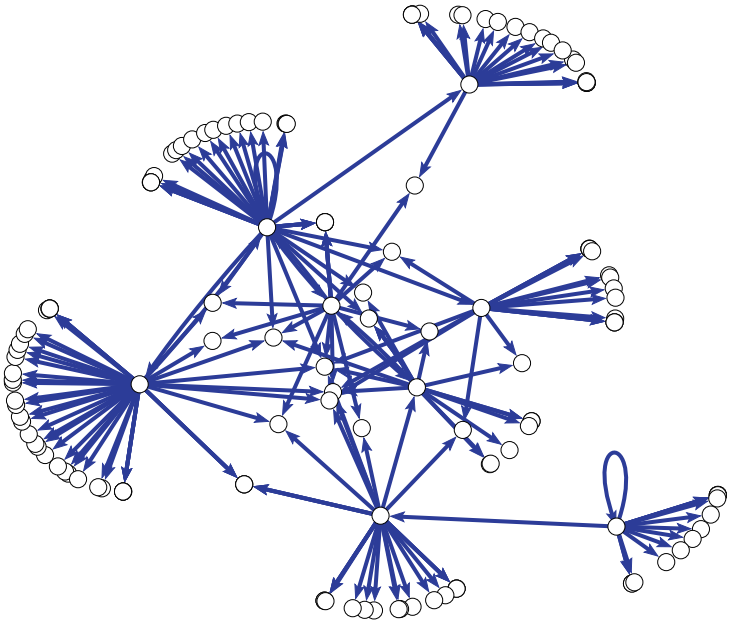

Fig. 1.4 AIDS Blog Network

The network in Figure 1.4 is a snapshot of the pattern of citation among 146 unique blogs related to AIDS, patients, and their support networks, collected by Gopal [184] over a randomly selected three-day period in August 2005. A directed edge from one blog to another indicates that the former has a link to the latter in their web page (more specifically, the former refers to the latter in their so-called 'blogroll'). Collection of such data is facilitated by the very Internet, protocols, and software that make blogging possible in the first place. The resulting measurements can include not only indications of pairwise blogger interactions, but also information on blog content. Interesting questions relate, for example, to the dynamics of formation of blog communities, the role(s) of individual or subgroups of bloggers in those communities, and the spread of information throughout a blog network.

We will encounter topics relevant to information networks like this in Chapter 3 (i.e., mapping networks), Chapter 4 (i.e., description of structure and patterns), Chapter 5 (i.e., network graph sampling), and Chapter 7 (i.e., link prediction).

1.3 About this Book

Given the vast array of contexts in which networks arise, the multitude of questions asked in regards to those networks, and the variety of measurements taken towards the goal of answering these questions, how can one propose to write a single book on general topics relevant to the statistical analysis of network data? The answer to this question lies in the fact that it is possible – and indeed quite useful – to categorize many of the various tasks faced in the analysis of network data in different domains according to a statistical taxonomy. In so doing, a certain order and structure emerges for presenting some core of the multitude of statistical methods and models proposed in this area by researchers across diverse disciplines. It is along the lines of such a taxonomy, with the goal of presenting such a core, that this book is organized.

Broadly speaking, the material in this book is broken down into topics of (i) descriptive methods, in Chapters 3 - 4, and (ii) modeling and inference, in Chapters 6 - 10, with Chapter 5 playing somewhat of a transition role between the two. Interwoven throughout, we integrate relevant issues of data collection, data management, and computing. Of course, in reality 'descriptive' and 'inferential' techniques are not always cleanly separated, nor in practice does the task of statistical analysis flow linearly from the former to the latter and simply stop, but rather is generally iterative. Nevertheless, we find this organization convenient.

In more detail, the contents of this book are as follows. Chapter 2 contains preliminary technical material on graphs, probability, and statistical inference necessary for the rest of the book. Following these preliminaries, methods for the descriptive analysis of network data are developed in Chapters 3 and 4. In the former we concentrate on the task of converting network measurements into a network graph representation, while in the latter we focus on the description of structure and the identification of patterns in such network graphs. In Chapter 5 we explore the effects of sampling on the extent to which characteristics of an observed network graph reflect the corresponding characteristics of the underlying network being studied. Then in Chapters 6 and 7 we turn to the task of modeling network graphs. In Chapter 6 we study models for describing an observed network graph, while in Chapter 7 we consider the problem of inferring a network graph based on incomplete or indirect measurements. Next, in Chapters 8 and 9 we turn to problems involving the modeling and inference of processes on a network graph. Chapter 8 concerns network-indexed processes, of both a static and dynamic nature, while Chapter 9 is devoted to the special case of network flow processes.

Finally, in Chapter 10 we briefly discuss the topic of graphical models. The graphical modeling paradigm differs from that associated with most of the models described in this book – wherein graphs serve either as data objects themselves or effectively as indexing for other data objects – in that graphs are used to describe the conceptual structure (or, more formally, collections of conditional independence relations) associated with statistical inference. Nevertheless, the two perspectives are by no means entirely distinct, and in fact we will encounter a number of instances of graphical models earlier in Chapters 6 through 9. The purpose of Chapter 10 is to

make more precise the role that can be played by graphical models in the analysis of network data.

We have aimed in the writing of each chapter to concentrate on what we see as a core set of topics relevant to the theme of the chapter. In doing so, we inevitably exclude certain worthwhile and valuable aspects of the literature. For example, most of the material in the book is developed around the case of static network graphs. Material on dynamic network graphs (i.e., those evolving in time), an increasingly active but less mature area of network research, is included where relevant in certain chapters as only a small subsection near the end. In an attempt to compensate somewhat for this and similar truncation or exclusion of topics, we have included additional references to the literature in the section 'Additional Related Topics and Reading' appearing at the end of each chapter. However, admittedly, even through this device it is impossible to be comprehensive.

This book is intended for students and researchers across a wide range of quantitative disciplines, including bioinformatics, computer science, economics, information science, mathematics, physics, sociometrics, and – of course – statistics. At a minimum, a reader will need a solid foundation in calculus and linear algebra, as well as the equivalent of a strong introductory course in probability and in statistical modeling. Readers with additional background will (hopefully!) reap additional benefit accordingly. For those readers outside of statistics who would like to establish a more thorough grounding in statistical modeling and inference beyond the level of an introductory course, we recommend Wasserman's *All of Statistics: A Concise Course in Statistical Inference* [392], a text written expressly for such readers.

The material in this book is generally presented in a manner that attempts to strike a balance between concepts and mathematical detail. It is expected that readers will want to follow the various threads woven throughout to their origins in the literature, and copious use of references has been made for this purpose. In addition, in order to encourage readers to explore certain topics in greater depth, we have included with each chapter a handful of exercises. These are a combination of analytical and computational exercises. The analytical exercises are often completion problems, picking up loose threads of a more technical nature from the main body of the chapter. The computational exercises are often open-ended, a format we have found useful in teaching this material to students of diverse backgrounds, allowing them to define and attack problems in specific topic areas in a manner most beneficial to whatever network-oriented agendas they bring to the course.

Lastly, we comment on the topic of software. The sheer variety of software available in this area for statistical analysis of network data mirrors the number of communities involved in such work. In this book we have chosen to be 'software agnostic,' in that we do not advocate nor solely use any one particular software for the examples presented. There are two reasons for this choice. First, no single software package at this time is best suited for conducting all of the statistical analyses described in this book – and it is not our aim to develop such a package. Second, when we have taught this material, our experience has been that students often come to the course already with a preferred package(s). Most of the network figures in

this book were created using `Pajek` [24, 111], while computing for the numerical illustrations generally was done using either the **R** [1] or **MATLAB**® software environments. We will attempt to maintain on the website for this book

http://math.bu.edu/people/kolaczyk/SAND.html

a (surely incomplete!) list of popular packages and software for network analysis. Readers are encouraged to help us maintain this list and to let us know of important omissions. Machine-readable copies of most data used in this book are also available from this website.

http://www.springer.com/978-0-387-88145-4

Statistical Analysis of Network Data
Methods and Models
Kolaczyk, E.D.
2009, XII, 386 p., Hardcover
ISBN: 978-0-387-88145-4