Toward a Probabilistic Foundation for Statistical Network Analysis

Harry Crane*

March 9, 2020

Abstract

This introductory chapter of *Probabilistic Foundations of Statistical Network Analysis* explains the major shortcomings of prevailing efforts in statistical analysis of networks and other kinds of complex data, and why there is a need for a new way to conceive of and understand data arising from complex systems.

"Networks are everywhere"

In recent years there has been an explosion of network data *— that is, measurements that are either of or from a system conceptualized as a network — from seemingly all corners of science.* (Kolaczyk [106])

Empirical studies and theoretical modeling of networks have been the subject of a large body of recent research in statistical physics and applied mathematics. (Newman and Girvan [83])

Networks have in recent years emerged as an invaluable tool for describing and quantifying complex systems in many branches of science. (Clauset, Moore and Newman [38])

Prompted by the increasing interest in networks in many fields [...]. (Bickel and Chen [19])

Networks are fast becoming part of the modern statistical landscape. (Wolfe and Olhede [155])

The rapid increase in the availability and importance of network data [...]. (Caron and Fox [32])

Network analysis is becoming one of the most active research areas in statistics. (Gao, Lu and Zhou [79])

Networks are ubiquitous in science. (Fienberg [74])

*Rutgers University, Department of Statistics & Biostatistics

> *Networks are ubiquitous in science and have become a focal point for discussion in everyday life.* (Goldenberg, Zheng, Fienberg, and Airoldi [84])

There is currently no shortage of interest in 'network science', 'network data', 'complex networks', or just about anything else that invokes the term 'network'; see, e.g., recent popular books on the topic [13, 151]. In writing this book, I have done my part in furthering this trend; and in reading it, so have you. But as it was never my intention to become part of the networks hype—a hype reflected in the quotes at the top of this page—I do not set out here to celebrate the importance of network science or its great 'successes' in better understanding the complexities of our world. To the contrary, while I acknowledge the potential of network science for gaining better insights about complex data structures and the systems that produce them, I also recognize that this potential has not yet been realized. Especially within statistics, the study of 'networks' has been greatly limited by a lack of appreciation for the complexity of 'network data' and a lack of creativity in developing new ways to think about those complexities. By now these limitations are woven so deeply into the fabric of statistical thinking that overcoming them is easier done by starting a new fabric, rather than modifying the existing one. So, in addition to clarifying the current limitations of statistical network analysis (in Chapters 1–4 and 6–7), I set out here along a new path with the hope of catching a glimpse of what lies ahead. And while certain parts of this book (e.g., Chapters 5, 8–11) do represent substantial progress in this direction, I make no claim to overcome all of these limitations here.

With these objectives in mind, this book is not intended as a survey of existing models or a catalog of currently available techniques for analyzing network data. The book is instead a *perspective* on how to better represent, model, and think about complex, heterogeneous data structures that arise in modern applications. The current ways of doing things, and their various extensions, are insufficient for this purpose. I discuss some early attempts at gaining such a new perspective throughout Chapters 7–11, but surely the future of statistical network analysis lies almost entirely beyond these pages, in a yet-to-be-celebrated breakthrough.

In venturing beyond the conventional graph-theoretic representation of networks and its associated random graph models, I am confident that the later chapters are a step in the right direction. But just as it is wrongheaded to believe that the current graph-theoretic convention is the 'correct', 'best', or 'only' way to think about network data, it would be foolish to suggest that any of these new approaches is absolutely superior to more conventional methods. To be sure, there are ways in which these new approaches provide a better perspective on network data of a certain kind. For example, the perspective of edge exchangeability (Chapter 9) allows us to express and extract properties from interaction data that standard vertex-centric approaches cannot. Such an expansion of the prevailing mindset, regardless of whether it proves 'useful' in any practical domain, is necessary to broaden the scope of statistical thinking beyond the traditional paradigm. Continued sharpening of perspective and enrichment of mindset, far beyond what came before and what lies within these pages, motivates everything that follows.

1 *Analogy*: Bernoulli trials

Network analysis is no more about studying Facebook, or Twitter, or the loyalties of karate club members [161] than classical statistics is about tossing coins. And yet, the

theory of coin tossing, as formalized by infinite sequences of independent, identically distributed (i.i.d.) Bernoulli trials, lays the groundwork for much of classical statistical theory; see, e.g., [71]. For an analogy, coin tossing is to the statistical analysis of simple, unstructured data as networks are to the statistical analysis of complex, dependent data:

$$\text{coin tossing} \ : \ \text{unstructured data} \ :: \ \text{network analysis} \ : \ \text{complex, structured data.}$$

From this analogy, I make a few initial observations.

First, just as i.i.d. Bernoulli trials are an entry point into classical statistics, through the law of large numbers, central limit theorem, etc., so too is network modeling an entry point into modern complex data analysis. Much like the classical theory of statistical inference is erected on the scaffolding of the i.i.d. sequence, the modern theory of inference from complex data will be built on the probabilistic foundations of statistical network analysis.

Second, instead of heralding the ubiquity of 'networks', as in the opening quotations, we would be better off recognizing the emergence of *complexity* in modern data science, where 'complexity' is used here to mean dependence, structure, heterogeneity, and the like. At present, networks are the primary vehicle for representing complex data structures and network analysis is the predominant method for understanding complexity, dependence, and heterogeneity.

Third, given the ubiquity of complexity and its many forms, statisticians can no longer rely on a limited toolbox of classical techniques and old ideas. New foundations for the statistical analysis of complex data must be forged; and these foundations cannot be derivative on the classical theory of linear models, i.i.d. sequences, etc. The newness of modern networks problems is paradigm-shifting, and thus warrants a shift in the paradigm within which we think about, discuss, and analyze such data. I clarify this point of view in the coming several pages, with special focus on the statistical foundations of network analysis, where they currently stand and where they are headed.

Probabilistic Foundations of Statistical Network Analysis emphasizes *modeling* (as a verb, the *act* of specifying a model), not *models* (the noun, those models which already exist). The reasons are manifold:

- One, the *act* of modeling should be thought of as an act of *imposing* structure on the data (and thus on the world). One does not simply *choose* a model from an existing class of acceptable choices. One instead *posits* a model, and in doing so declares how the data behaves and how that behavior fits into a bigger picture. Classical statistics, which deals primarily with data having little or no internal structure (i.e., sequences and sets), has conditioned the statistician to behave rather lazily when choosing a model. Since there is little structure in many classical datasets, the act of modeling involves little more than identifying a family of probability distributions to describe a (nearly) structureless collection of measurements. (To be clear, I am not claiming that classical data sets lack structure; rather, I am observing that their conventional representation, most often as sets of points in $\mathbb{R}^d$, and the models chosen to describe them, e.g., often i.i.d. or exchangeable models, tend to minimize the impact of this structure on data analysis.) When dealing with structured data—and in the case of network data, the structure *is* the data—the act of imposing structure (via modeling) should be taken much more seriously.

- Two, most of the network models that already exist are inadequate for modern network data structures. They do not live up to their name as 'models' in the vast

majority of situations. We encounter several examples throughout Chapters 2 and 6–8.

- Three, even though existing network models (i.e., stochastic blockmodels, exponential random graph models, graphons) are known to suffer serious drawbacks for modern applications, their appearance throughout the theoretical and applied literature remains pronounced. I have no desire to continue this trend.

- Four, a major reason for the continued use of these limited models seems to be a general lack of interest in positing new ones. The canonical statistics curriculum focuses primarily on the analysis and application of standard models (Binomial, Poisson, Gaussian, Exponential) but without emphasizing the principles that make these models 'good' in any given situation. Rather than fret over the technicalities and nuances of constructing better models, students and researchers are instead indoctrinated with the Boxian trope, "All models are wrong, but some are useful" [26], without any clarity as to why models are 'wrong' or what makes them 'useful'. With Box's proverb comes the demotion of models and modeling, and the elevation of estimation, prediction, approximation, and computation.

Perhaps the Boxian proverb does little harm in the classical paradigm, where laws of large numbers, the central limit theorem, and asymptotic approximations abound. But it is untenable within the emerging paradigm of network analysis, in which there are few reliable asymptotic results; and those asymptotic results that do exist are hard to make sense of, e.g., minimax rates for graphon models, consistency properties for stochastic blockmodels and exponential random graph models, and asymptotic sparsity properties of so-called 'sparse graphon' models (Section 7.2). Bear in mind: the model is what the researcher puts in. Everything else is either given (i.e., data) or derived (i.e., inferred). The choices made while modeling—how one chooses to 'look at' and 'think about' the data—are most critical to determining whether the resulting inferences are 'useful', in Box's parlance. As I emphasize with the statistical modeling paradigm of Chapter 5, whether the result of an analysis 'is useful' or 'makes sense' or 'is valid' cannot be assessed solely on whether the estimators are unbiased, consistent, efficient, etc., as these diagnostics are meaningless unless grounded by an internally coherent model. No matter how much statistical inference is presented as an 'objective' approach to data analysis, modeling is undoubtedly a subjective and personal activity. And so it ought to be taken personally.

With the discussion below, I hope more than anything else to restore modeling to its role at the center of the statistical paradigm, bridging the divide between data collection and inference. Along the way I will carefully consider Box's admonition—to employ models that are 'useful'—along with other foundational topics (i.e., symmetry and exchangeability) at the heart of statistical inference. For the most part, I have chosen to deemphasize technical aspects of network analysis in favor of high-level concepts, both in the remainder of this chapter and throughout the book. For the rest of this opening chapter, I discuss the guiding principles of statistical network analysis at a high level. Although the technical aspects of this chapter are light, the concepts are subtle, and are essential in order to appreciate the core ideas motivating this book.

2 *What it is*: Graphs vs. Networks

In these pages, the term 'network' refers to a specific instantiation of what can be vaguely understood as 'complex data'. But even in the specific case of 'network data', it is important to distinguish between the fundamental objects of study (i.e., networks) and the conventional mathematical representation of those objects (i.e., graphs). The distinction between networks and graphs marks the initial divergence between the perspective put forward here and the prevailing 'networks-as-graphs' perspective found throughout the literature.

To be clear: *networks are not graphs*. A *graph* is a mathematical object consisting of a set of vertices V and a set of edges $E \subseteq V \times V$. This mathematical concept can be extended in several ways to allow for multiple edges, hyperedges, and multiple layers, but all of these objects, i.e., graphs, multigraphs, hypergraphs, multilayer graphs, etc., are mathematical entities. They are also distilled entities, in that they can be discussed independently of any presumed statistical or scientific context. From this point of view, graphs can be regarded as a 'syntax' for communicating about network data. But this graph-theoretic syntax is just one language with which to communicate about networks. And like any language, it is limited in what it can express. In becoming too attached to this one language for talking about networks, we limit the nature of insights that can be gleaned from network data. A sure sign of progress in the foundations of network analysis is the development of new ways to express and understand network data. See Chapters 9 and 10 for one such new approach.

A *network*, on the other hand, is an abstract concept referring to a system of interrelated entities. For us, the concept of 'network' is neither concrete nor well-defined, but rather is vague and amorphous, emerging from an intuitive judgment about perceived structure in an observed system. For example, import-export partnerships between countries, social relationships among high school students, patterns in phone call activity, connectivity among Internet servers, and interactions among genes all invoke the concept of a *network* of relationships or interactions in a particular context. Although it is sometimes reasonable to represent these networks mathematically as graphs, the systems are not graphs in themselves. For example, the Internet is a physical structure consisting of wires, servers, and routers. A graph is a set V together with another set $E \subseteq V \times V$. The physical Internet invokes the concept of a 'network', and some aspects of it can be represented or modeled as a graph, but the Internet is not a graph.

Moving beyond graphs

The reader who has read the word 'network' and every time envisioned a 'graph' faces a steep *unlearning* curve to appreciate the richness of structure encoded in the concept of 'network'. If there is to be progress in understanding complex, structured data, then the conventional ways of thinking about ordinary, unstructured data—the data sequences and arrays that fill statistics textbooks—must be purged from memory, or at least demoted from their default status in data analysis. To think about networks properly, one must strongly resist any temptation to embed networks in Euclidean space, or use the terms 'network' and 'graph' interchangeably, or any similar such urge to impose the flat view of data taken by classical statistics on the voluminous and rich structure which the concept of network calls into being.

Though I strongly advocate this point, it is with great regret that almost all of the

'networks' discussed in this book are treated as ordinary 'graphs', an exception being the important class of edge and relationally exchangeable network models in Chapters 9 and 10. This antithetical presentation can be explained by the extraordinary primitiveness in the current state of affairs. The concept of 'network data' is itself a very special case—the base case—of what can be understood as 'complex data'. The mathematical language of graph theory studies the even more restricted class of 'networks' which can be represented as pairs (V, E) consisting of a set V of vertices and a set $E \subseteq V \times V$ of edges. The recent proposal of edge-labeled networks (Chapter 9) breaks free of this traditional view and inspires hope for expanding the scope of 'network analysis' beyond what is currently conceivable, but there is still a long way to go.

3 *How to look at it*: Labeling and representation

Think of statistical analysis as the act of discerning the nature of some large, complex object in a dark room. You only have a flashlight, which can illuminate just a small piece of that object. In this analogy, the illuminated piece is the data on which your inference about the large, unobservable object is to be based. Different angles of shining the flashlight can be understood as different ways of looking at, or representing, the data. For example, the representation of a network as a vertex-labeled graph (Figure 1(b)) corresponds to the shadow cast by shining the light from one angle; the edge-labeled graph (Figure 1(c)) is the shadow cast from a different angle. Both are shadows of the same object, namely Figure 1(a), but the angle from which the light is shone (i.e., the perspective from which the data is viewed) affects which attributes are visible and which are obscured, and thus which inferences the data supports and which it does not.

Because in many classical applications there is just one canonical angle from which to look at the data, it is easy to overlook the role played by 'perspective' in complex data analysis. In a sequence, for example, the measurements $X_1, X_2, \ldots$ contain the primary information. Changing the 'angle' from which we shine our proverbial flashlight on this data (e.g., by converting pounds to kilograms, or feet to inches) does not change the nature of the measurements $X_1, X_2, \ldots$. But the significance of this 'angle' cannot be overstated when handling networks and other complex data structures. In these latter instances, the structure *is* the data, and different aspects of this structure may be visible depending on the angle from which the light is shone.

In practice, this 'angle' is manifested first and foremost in how the network is represented, for which the choice of labeling is a basic consideration. In Figure 1, for example, the 'unlabeled' structure in Figure 1(a) is the object of interest. Ideally, we would treat this 'unlabeled' structure as the data and analyze it directly, but this is not possible. Unlabeled structures cannot be treated as data because unlabeled objects cannot be *represented*. To analyze data one must be able to talk about it; and to be able to talk about something, one must assign names to whatever parts of that thing are being discussed. For networks, this 'naming' comes in the form of labeling the constituent parts of the data. Without such a labeling, we cannot even utter a word.

To make this point clear, realize that the object in Figure 1(a) merely *represents* the abstract notion of an unlabeled network. But the object itself is labeled by the spatial orientation of its edges and vertices on the page. This spatial orientation allows one to speak (i.e., 'utter') about this network by referring to the relative positions of vertices/edges, e.g., by pointing or describing the positions in words. Mathematically,

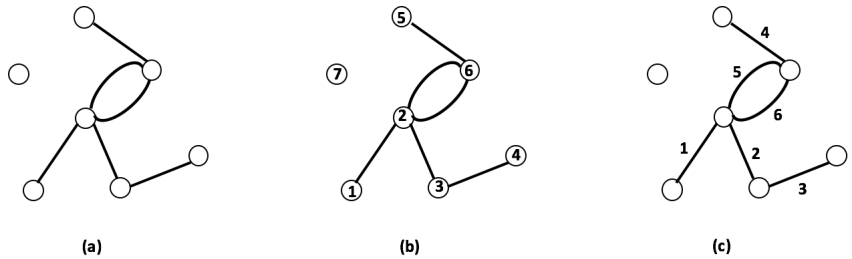

Figure 1: Two perspectives on network data. (a) Represents the essential structure of 'unlabeled' network data. (b) Represents the structure in (a) by assigning labels to its vertices (i.e., vertex-centric perspective). (c) Represents the structure in (a) by assigning labels to its edges (i.e., edge-centric perspective).

such 'unlabeled' structures are typically represented by 'removing the labels' and working with a class of structures that are equivalent up to relabeling. But the appropriate notion of equivalence itself depends on the perspective from which two networks are to be treated as 'equivalent'. For example, the equivalence class of vertex-labeled networks (as in Figure 1(b)) differs from the equivalence class of edge-labeled networks (as in Figure 1(c)), because the corresponding notions of equivalence differ based on the chosen perspective. Which perspective is appropriate for a given application depends on the context.

4 *Where it comes from*: Context

Given the diverse scenarios in which networks arise, there can be no single 'correct' approach to network analysis. Instead, what makes 'network analysis' relevant to a given problem depends on the *context*. And this context should be accounted for at every stage of the analysis, beginning with the way in which the data is represented, continuing through model specification, and culminating in inference. As emphasized in the previous section: the representation of network data reflects the perspective from which it is being analyzed, which in turn determines what inferences can be drawn from the analysis. To elicit the best available insights from the data, we want to shine our flashlight (i.e., represent and model the data) from the optimal angle, and the optimal angle in any given application depends crucially on the context.

Consider the structures in Figure 1. Do they represent the same network? Perhaps. Assuming they do represent the same network, do they provide the same *representation* of that network? Of course not. Figure 1(a) represents the 'shape' of the network, without explicitly identifying any of its other components, e.g., vertices or edges. Figure 1(b) identifies each vertex with a distinct label. Figure 1(c) identifies each edge with a distinct label, leaving vertices unlabeled. But what's the difference? The difference, we will see throughout the coming chapters, is a matter of perspective. In labeling the vertices, Figure 1(b) asserts a 'vertex-centric' view of the shape in Figure 1(a), and this vertex-centric view differs from the 'edge-centric' point of view taken in Figure 1(c). Even though these may be different representations of the same network, the choice of representation reflects the perspective of the data analyst and the context of the application, both of which affect inference.

5 *Making sense of it all*: Coherence

There are primarily two aspects to network modeling. The model first *describes* the observed data from the perspective of the statistician. And then, to draw inferences beyond the observed data, the model specifies a *context* in which to interpret the data. With this, the model has two components:

- a *descriptive* component consisting of the family of candidate probability distributions for describing variability in the observed data, and
- an *inferential* component explaining how the observed data fits into a larger context.

Both components are essential to proper model specification and sound statistical inference.

Returning to the Boxian proverb, "All models are wrong, but some are useful," I regard 'making sense' as the first step towards 'being useful'. To make sense, the inferences based on the model should be interpretable within a single (coherent) context. This observation culminates in the formal concept of *coherence*, by which the *description* of the model 'fits coherently' into its *context* in a sense made precise in Definitions **??**–**??**. (See Chapter 5 for a more formal discussion of coherence and its significance for statistical inference.)

Beyond coherence, there are often practical considerations regarding whether or not the presumed context is suitable, or whether the specified model can actually be used (i.e., computed) in a given application. But such practical matters should be considered only after minimal logical conditions, such as coherence, are met. Without coherence, any computational or practical techniques which enhance the analysis are of little use, precisely because the model which they will have enhanced does not make sense.

6 *What we're talking about*: Examples of network data

Throughout these pages, we will encounter a number of scenarios under which different modeling considerations are appropriate. Whenever possible, I try to motivate these scenarios by real (or realistic) applications for which canonical examples already exist. I survey some of these common scenarios below. For the most part, these examples are not interesting on their own, and are offered here only to illustrate how basic principles of network analysis arise in practice.

6.1 Internet

Several early developments in network science grew out of empirical observations taken on the Internet, defined as the network of servers and systems connected by physical wires. Of all the datasets discussed here, the Internet is one of the only 'real' networks in the sense that it corresponds to an actual physical object. Explaining the observed power law structure in sampled data from the Internet and World Wide Web was one of the primary motivations for Barabási and Albert's preferential attachment model. The widespread empirical observations of power law degree distribution, both in the Internet and other real-world networks, remains one of the most evocative illustrations of the effects of sampling on network analysis, which have been mostly overlooked until recently [52, 54, 112, 127, 154]. I discuss the role of sampling further in Chapter 3.

Because of its physical nature, the Internet network invokes a notion of 'ground truth' that is absent from other familiar applications in network science. For example, community detection in social networks seeks an optimal clustering of vertices into (disjoint) communities based on their network connectivity. As the concept of 'social network' is itself a nebulous one, in many cases there is no 'true' division of vertices against which to assess the inferred clustering. (A notable exception is the karate club network of Zachary [161], see Section 6.3. But in modern network analysis, the karate club network is treated more as a meme than as a serious dataset.[1])

6.2 Social networks

In social network analysis, vertices represent individuals and edges represent social ties between their adjacent vertices. The network does not correspond to anything physical, as in the Internet, but rather represents invisible social forces driving interactions within a population, e.g., shared recreational interests, common political views, or professional relationships. I discuss some scenarios of social network modeling in Chapters 2, 3, 6, 7, and 8.

6.3 Karate club

The karate club dataset [161] records social interactions among 34 members of a university karate club for the three-year period spanning 1970–1972. Represented as a network with multiple edges, each vertex corresponds to a different member of the club and each edge corresponds to a different social interaction between the corresponding club members. Since all club members have been observed, the dataset exhibits no vertex sampling or growth. Zachary's initial analysis highlighted the division of members into two factions, caused by a rift between the club's two leaders. This known separation of its vertices makes the karate club network a canonical testbed for community detection methods. The standard analyses of the karate club also demonstrate a common pitfall of network analysis, which I discuss further in Chapter 9.

6.4 Enron email corpus

The Enron email corpus [104] consists of email activity for 150 employees at the Enron corporation. The dataset contains not only information about the senders and recipients of emails but also textual content, timestamps, etc.[2] Most relevant for our purposes is the network structure induced by the exchange of emails between employees, which we construct by letting each edge correspond to a different email in the corpus. An important difference from the karate club network (Section 6.3) is that a single edge (i.e., email) can involve more than two vertices (i.e., sender/recipient). For example, an email sent from employee A to employees B, C, and D corresponds to a single (hyper)edge in the network representation. Interaction networks such as this and the collaboration networks discussed next are the subject of Chapters 9 and 10.

[1] Since 2013, the 'Zachary Karate Club Club' (ZFCC) trophy has been presented, as a joke, at various conferences to the speaker who first mentions the karate club network in his or her presentation. See http://networkkarate.tumblr.com/ for more information.

[2] See http://www.cs.cmu.edu/~enron/ and [131] for some applications involving this dataset.

6.5 Collaboration networks

Collaboration networks between actors [104, 134], scientists, authors, and other communities of professionals have much in common with the above Enron dataset. In an actors network, for example, each actor corresponds to a different vertex and each movie corresponds to a different edge consisting of the set of all vertices whose associated actors play a role in that movie. A common feature of the karate club, Enron, and collaboration networks is their growth by sequential addition of new edges, in the form of interactions, as opposed to sequential addition of new vertices, as in the preferential attachment model (Chapter 4). This feature of interaction networks figures prominently in Crane and Dempsey's framework of edge exchangeability (see [54] and Chapters 9–10).

6.6 Blockchain and cryptocurrency networks

Cryptocurrencies, such as Bitcoin [10, 122], Ethereum [29], and RChain [40], combine several innovative ideas in an effort to revolutionize economic activity through the use of peer-to-peer networks, blockchain technology, and smart contracts. These 'digital currencies', e.g., Bitcoin, operate on a blockchain, which records all transactions in a 'ledger' that stores the complete history of all Bitcoin transactions. This ledger is maintained by a distributed peer-to-peer network, which updates the blockchain by adding blocks according to a majority voting consensus protocol. Peer-to-peer networks also play an important role in decentralizing control of the network away from a centralized authority toward a distributed collection of nodes in the network. All of these components come together to create a complex network of transactions between addresses on the blockchain. As this revolutionary new technology matures, blockchain data should serve as a fertile testbed for model development at the frontiers of complex data analysis.

6.7 Other networks

In addition to the above examples, there are networks from social media platforms such as Facebook and Twitter [117], brain networks [82], gene regulatory networks, telecommunications networks, wireless sensor networks [100, 101], etc. All of these are just a small selection of the many structures that are now referred to as 'network data'. Because I focus in this book on establishing the foundations of network analysis, I do not undertake any detailed application of a particular network dataset. These examples do, however, provide concrete modeling 'scenarios' within which to discuss different modeling approaches. The 'scenarios' accompanying each new class of models are meant to provide additional context for the more technical aspects of network analysis covered throughout the text.

6.8 Some common scenarios

As the scope of networks expands to encompass problems in new disciplines, so too must the mathematical and statistical techniques available to address these problems. I conclude this section with a brief review of some of the basic contexts for network modeling in social science, epidemiology, and national security. In the near future, it seems inevitable that the relevance of networks will continue to expand to include a wider range of disciplines as human behaviors and complex systems become ever more entangled

through the growth of the Internet, social media, and other emergent technologies, such as blockchain.

Social science. Social network analysis was the primary domain of statistical network analysis until the mid-1990s. By all known accounts, the study of social networks began with Moreno's invention of the sociogram in 1930 [121]. Still today, many common network models (e.g., stochastic blockmodels (SBMs) [89] and exponential random graph models (ERGMs) [78, 90]) were originally motivated by sociological applications. With the growth of online communities and social media as a way to consume and disseminate information, traditional social networks have given way to networks with much more complex structure than traditional social network models, namely SBMs and ERGMs, are equipped to handle.

Epidemiology. Stochastic process models for disease spread on networks garner substantial interest in applied probability and statistical physics. The now classical SI (susceptible-infected), SIS (susceptible-infected-susceptible), and SIR (susceptible-infected-recovered) models describe how infections spread in a population whose interactions are represented by a graph. In the SIR model, for example, each node fluctuates among three states: susceptible to infection (S), infected (I), or recovered (R). As time evolves, infected individuals randomly transmit the disease to their susceptible neighbors. Infected individuals recover, and are henceforth immune from infection, according to another random process. Basic questions center around how the different combinations of network structure and disease dynamics affect disease spread. For example, given certain initial conditions, what is the probability of an epidemic, i.e., the disease spreads to a non-negligible fraction of the population? One can also imagine how such models could be useful for designing effective advertising strategies or for modeling how information percolates through social networks.

National security. Networks arise in national security in at least two different ways. There are physical networks, such as the Internet, the U.S. Power Grid, and the transportation network of roads, bridges, and highways, all of which must be protected against failure or targeted attack. In fact, many experts [37] now regard cyberspace as the primary battlefield of modern warfare and national security, making resilience to targeted network attacks critical to national security interests. These concerns over cybersecurity, and the role of network science in resolving them, will continue to grow as more economic and social activity transitions to cyberspace.

Non-physical networks also play a role in national security, as terrorist organizations rely on complex webs of social, financial, and political interactions in order to evade detection [135]. As long as critical national infrastructure is controlled by a centralized, bureaucratic government, the governed society is vulnerable to both external attacks (e.g., hacking) and internal attacks (e.g., leaks), both of which have become increasingly prevalent and widely publicized in recent times. As a countermeasure to the vulnerability and antiquity of centralized authority, blockchain technology and cryptocurrencies (Section 6.6) distribute control of currency and other critical information to a "trustless" peer-to-peer network [10, 122]. The use of networks for this purpose is likely to have potential national security implications moving forward.

7 Major open questions

The probabilistic foundations of statistical network analysis currently face a few major open questions that are worth keeping in mind over the coming chapters.

7.1 Sparsity

Early interest in network science grew out of several concurrent empirical observations of sparsity and so-called 'scale-free' structure in real-world networks [1,5,14,70,111,113]. (Refer to Chapters 4, 7, and 9 for a more detailed discussion of sparsity and power law, i.e., scale-free, properties.) For present purposes, it is sufficient to interpret 'sparsity' to mean that the network has few connections relative to its size. For example, when represented as a graph with n vertices, a network is sparse if it has a negligible number of edges compared to the number n^2 of all possible edges. In statistics, sparsity draws interest for two competing reasons, which together capture the tension between empirical properties of network data and logical principles of statistical modeling. First, many sparse networks are observed to be well-connected as a result of heterogeneous patterns of connectivity (e.g., 'scale-free' structure). So while the network is in one sense poorly connected (because it is sparse), it is at the same time well-connected (because of its complex patterns). Second, the prevailing approaches to network modeling (e.g., stochastic blockmodels, graphons, and exponential random graph models) are unable to account for these observed empirical behaviors. These competing elements of network modeling have stalled progress in statistical network analysis for nearly a decade, primarily due to unrecognized limitations of conventional approaches. Chapters 9–10 present one attempt to address this challenge, which interested readers are encouraged to build upon.

7.2 Modeling network complexity

In addition to sparsity, other heterogeneous features of real-world networks, such as power law degree distributions, clustering, and the 'small-world' property [152], confound attempts to analyze network data with standard models. In this opening chapter, I have emphasized the need for new tools to conceptualize the complexity of modern data structures. Above all, we seek to work with the complexity of network data, rather than fight against it by reducing complexity to something with less structure. This latter attitude of 'flattening' network structure is common throughout statistical analysis, and especially in network community detection, where non-overlapping subsets (i.e., communities) are sought to provide a 'low resolution' summary of much richer network structure. Community detection has become a cottage industry among statisticians interested in network analysis, but it is mostly counterproductive for understanding data complexity. I discuss models for community detection in the context of relative exchangeability (Chapter 8).

7.3 Sampling issues

Understanding the impact of sampling is one of the longest standing challenges in modern network science. Empirical observations of power law degree distribution in the Internet and other real-world networks [1, 5, 14, 70, 111] raise the question of whether these observed properties reflect the actual network structure or are merely an artifact of sampling bias [27, 112, 154]. This question is of central importance to statistical network analysis,

for which the mode of sampling establishes the essential link between observed and unobserved parts of the network needed for inference. But even as interest in network analysis has grown among statisticians, there has not been much effort to incorporate sampling into the theoretical foundations of the subject. Much of the work on network analysis promoted by flagship statistics journals consists of asymptotic results and standard analyses under models that are known to be inadequate for most serious applications (e.g., graphons, stochastic blockmodels, and exponential random graph models). Remarkably few of these analyses acknowledge the importance of sampling to network analysis; and those that do, e.g., [138], assume a stylized form of sampling by vertex selection which does not even remotely resemble the way in which real-world networks are sampled. I discuss these issues at length throughout Chapters 3–5, and again in Chapters 6 and 9.

7.4 Modeling network dynamics

While much of this book is dedicated to modeling single instances of a network, there is emerging interest in analyzing dynamic network data, such as temporal observations of brain activity and social media interactions. But so far statistical work on dynamic networks is mostly confined to theory and applications for the temporal exponential random graph model or other *ad hoc* approaches. Because network dynamics add another dimension to the already challenging problem of network modeling, the foundations of dynamic network analysis are even more technically and conceptually challenging than their non-dynamic counterpart. Chapter 11, in which I give a brief non-technical overview of some otherwise technical work from the stochastic processes literature [44,48,57], offers a potential starting point for a more general theory of dynamic network modeling. More in depth coverage of dynamic network analysis is beyond the scope of this book and is left as a topic worthy of its own book length treatment.

8 Toward a Probabilistic Foundation for Statistical Network Analysis

In this opening chapter I have laid out a vision for network analysis as the foundation for what I am calling *complex data analysis*. As of yet, this vision has not been realized, but it is my hope in this book to clarify the major tenets underlying this vision and, if possible, to light the path toward its ultimate fulfillment. If nothing else, I hope to convince readers that real progress in the analysis of complex data will be limited as long as the field continues to seek incremental advances within the networks-as-graphs orthodoxy. The ideas in Chapters 5 and 9–11 offer some first attempts to get beyond these limitations, but many challenges still lie ahead.

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook. A functional approach to external graph algorithms. *Proceedings of the 6th European Symposium on Algorithms*, pages 332–343, 1998.

[2] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the Bias of Traceroute Sampling or, Power-law Degree Distributions in Regular Graphs. *STOC '05*, 2005.

[3] N. Ackerman. Representations of Aut($\mathcal{M}$)-Invariant Measures: Part 1. *Accessed at arXiv:1509.0617*, 2015.

[4] N. Ahmed, J. Neville, and R. Kompella. Network Sampling: From Static to Streaming Graphs. *ACM Transactions on Knowledge Discovery from Data*, 8, 2014.

[5] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 171–180, New York, 2000. ACM Press.

[6] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[7] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Modern Phys.*, 74(1):47–97, 2002.

[8] D.J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[9] D.J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[10] A.M. Antonopoulos. *Mastering Bitcoin: Programming the Open Blockchain, 2nd Edition*. O'Reilly Media, 2017.

[11] A. Athreya, D.E. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D.L. Sussman, M. Tang, J.T. Vogelstein, and C.E. Priebe. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, in press, 2017.

[12] T. Austin. On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probability Surveys*, 5:80–145, 2008.

[13] A.-L. Barabási. *Linked: How Everything is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, 2003.

[14] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[15] S. Basu, A. Shojaie, and G. Michailidis. Network Granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453.

[16] J. Bertoin. *Random Fragmentation and Coagulation Processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[17] S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 803–812, 2008.

[18] S. Bhamidi, J.M. Steele, and T. Zaman. Twitter event networks and the superstar model. *Annals of Applied Probability*, 25(5):2462–2502, 2015.

[19] P. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21068–21073, 2009.

[20] B. Bloem-Reddy and P. Orbanz. Random walk models of network formation and sequential Monte Carlo methods for graphs. *Accessed at arXiv:1612.06404*, 2016.

[21] B. Bollobás. *Random Graphs, 2nd Edition*, volume 73 of *Cambridge Series in Mathematics*. Cambridge University Press, 2001.

[22] B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs. In *In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore)*, pages 132–139. New York, 2003.

[23] C. Borgs, J.T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *Accessed at arXiv:1601.07134*, 2016.

[24] C. Borgs, J.T. Chayes, H. Cohn, and V. Veitch. Sampling perspectives on sparse exchangeable graphs. *Accessed at arXiv:1708.03237*, 2017.

[25] C. Borgs, J.T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Accessed at arXiv:1401.2906*, 2014.

[26] G.E.P. Box and N.R. Draper. *Empirical Model Building and Response Surfaces*. John Wiley & Sons, New York, NY, 1987.

[27] A.D. Broido and A. Clauset. Scale-free networks are rare. *Accessed at https://arxiv.org/pdf/1801.03400.pdf on February 16, 2018*, 2018.

[28] C. J. Burke and M. Rosenblatt. A Markovian function of a Markov chain. *Ann. Math. Statist.*, 29:1112–1122, 1958.

[29] V. Buterin. Ethereum "white paper". *Accessed at https://github.com/ethereum/wiki/wiki/White-Paper on February 13, 2018*.

[30] D. Cai, T. Campbell, and T. Broderick. Edge-exchangeable graphs and sparsity. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4242–4250. Curran Associates, Inc., 2016 (appeared online November 20, 2016).

[31] F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. *Accessed at arXiv:1401.1137*, 2014.

[32] F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society, Series B*, 79(5), 2017.

[33] F. Caron and J. Rousseau. On sparsity and power-law properties of graphs based on exchangeable point processes. *Accessed at arXiv:1708.03120*, 2017.

[34] S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Annals of Statistics*, 41(5):2428–2461, 2013.

[35] D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

[36] F. Chung and L. Lu. *Complex Graphs and Networks*, volume 107 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2006.

[37] R.A. Clarke and R.K. Knake. *Cyber War: The Next Threat to National Security and What to Do About It*. HarperCollins, New York, 2010.

[38] A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[39] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[40] RChain Cooperative. *Accessed at https://medium.com/rchain-cooperative on February 13, 2018*.

[41] O.T. Courtney and G. Bianconi. Dense Power-law Networks and Simplicial Complexes. *Accessed at arXiv:1802.01465*, 2018.

[42] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman & Hall, London, 1974.

[43] H. Crane. The cut-and-paste process. *Annals of Probability*, 42(5):1952–1979, 2014.

[44] H. Crane. Time-varying network models. *Bernoulli*, 21(3):1670–1696, 2014.

[45] H. Crane. Rejoinder: The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1):37–39, 2016.

[46] H. Crane. The ubiquitous Ewens sampling formula (with discussion and a rejoinder by the author). *Statistical Science*, 31(1):1–39, 2016.

[47] H. Crane. Comment on F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society, Series B*, 79(5), 2017.

[48] H. Crane. Exchangeable graph-valued Feller processes. *Probability Theory and Related Fields*, 168(3–4):849–899, 2017.

[49] H. Crane. Randomly evolving graphs and their graph limits. *Annals of Applied Probability*, 26(2):691–721, 2017.

[50] H. Crane. Combinatorial Lévy processes. *Annals of Applied Probability*, 28(1):285–339, 2018.

[51] H. Crane and W. Dempsey. Community detection for interaction networks. *Accessed at arXiv:1509.09254*, 2015.

[52] H. Crane and W. Dempsey. A framework for statistical network modeling. *Accessed at arXiv:1509.08185*, 2015.

[53] H. Crane and W. Dempsey. Relational exchangeability. Accessed at arXiv:1607.06762, 2016.

[54] H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, in press, 2017.

[55] H. Crane and W. Dempsey. A framework for statistical network modeling. *First version, Accessed at arXiv:1509.08185v1*, September 28, 2015.

[56] H. Crane and S.P. Lalley. Convergence rates of Markov chains on spaces of partitions. *Electronic Journal of Probability*, 18(paper no. 61):1–23, 2013.

[57] H. Crane and H. Towsner. The structure of combinatorial Markov processes. *Accessed at arXiv:1603.05954*, 2016.

[58] H. Crane and H. Towsner. Relative exchangeability with equivalence relations. *Archive of Mathematical Logic*, in press, 2017.

[59] H. Crane and H. Towsner. Relatively exchangeable structures. *Journal of Symbolic Logic*, in press, 2017.

[60] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68.

[61] P. Diaconis and D. Freedman. On the statistics of vision: The Julesz conjecture. *J. Math. Psychol.*, pages 112–138, 1981.

[62] P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008.

[63] P. Doreian and F. N. Stokman eds. *Evolution of Social Networks*. Routledge, Mahway, NJ, 1997.

[64] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[65] D. Durante and D.B. Dunson. Nonparametric Bayes dynamic modelling of relational data. *Biometrika*, 101(4):883–898, 2014.

[66] D. Durante, N. Mukherjee, and R.C. Steorts. Bayesian Learning of Dynamic Multilayer Networks.

[67] R. Durrett. *Random Graph Dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2007.

[68] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[69] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112, 1972.

[70] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *ACM Comp. Comm. Review*, 29, 1999.

[71] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley, 1957.

[72] S. Feng. *The Poisson-Dirichlet Distribution and Related Topics*. Probability and its Applications. Springer-Verlag, Berlin, 2010.

[73] T. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1(2):209–230, 1973.

[74] S.E. Fienberg. A Brief History of Statistical Models for Network Analysis and Open Challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, 2012.

[75] O. Frank. Network sampling and model fitting. In *Models and Methods in Social Network Analysis*, pages 31–56. Cambridge University Press, New York, 2005.

[76] O. Frank. Estimation and sampling in social network analysis. In *Encyclopedia of Complexity and Systems Science*, pages 8213–8231. Springer, New York, 2009.

[77] O. Frank. Survey sampling in networks. In *Handbook of Social Network Analysis*. Sage, London, 2011.

[78] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

[79] C. Gao, Lu. Y., and H.H. Zhou. Rate optimal graphon estimation. *Annals of Statistics*, 43(6):2624–2652, 2015.

[80] F. Gao and A. van der Vaart. On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees. *Stochastic Processes and Their Applications*, 2017.

[81] E.N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[82] C.E. Ginestet, P. Balachandran, S. Rosenberg, and E.D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *Annals of Applied Statistics*, 11(2):725–750, 2017.

[83] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[84] A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):1–117, 2009.

[85] M.S. Handcock and K.J. Gile. Modeling social networks from sampled data. *Ann. Appl. Stat.*, 4(1):5–25, 2010.

[86] S. Hanneke, W. Fu, and E.P. Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

[87] S. Hanneke and E.P. Xing. Discrete temporal models of social networks. In In E. Airoldi, David M. Blei, S.E. Fienberg, A. Goldenberg, E.P. Xing, and A.X. Zheng, eds.*, Statistical Network Analysis: Models, Issues, and New Directions: ICML 2006 Workshop on Statistical Network Analysis*, volume 4503 of *Lecture Notes in Computer Science*, pages 115–125. Springer, 2007.

[88] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, 97(460):1090–1098, 2002.

[89] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[90] P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, pages 33–65, 1981.

[91] P. Holme and J. Saramäki (eds.). *Temporal Networks*. Understanding Complex Systems. Springer, 2013.

[92] D.N. Hoover. Relations on Probability Spaces and Arrays of Random Variables. Preprint, Institute for Advanced Studies, 1979.

[93] D.R. Hunter, S.M. Goodreau, and M.S. Handcock. Goodness of Fit of Social Network Models. *Journal of the American Statistical Association*, 2008.

[94] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

[95] S. Janson. On edge exchangeable random graphs. *Accessed at arXiv:1702.06396*, 2017.

[96] P. Ji and J. Jin. Coauthorship and citation networks for statisticians (with discussion and rejoinder by the authors). *Annals of Applied Statistics*, 10(4):1779–1812, 2016.

[97] O. Kallenberg. Exchangeable random measures in the plane. *Journal of Theoretical Probability*, 3(1):81–136, 1990.

[98] O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. Springer, 2005.

[99] B. Karrer and M.E.J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011.

[100] N. Katenka, E. Levina, and G. Michailidis. Local vote decision fusion for target detection in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(1):329–338.

[101] N. Katenka, E. Levina, and G. Michailidis. Detection, Localization, and Tracking of a Single and Multiple Targets with Wireless Sensor Networks. *Computational Network Theory: Theoretical Foundations and Applications*, 5, 2015.

[102] M. Khabbazian, B. Hanlon, Z. Russek, and K. Rohe. Novel Sampling Design for Respondent-driven Sampling. *Accessed at arXiv:1606:00387*, 2016.

[103] M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, and M.A. Porter. Multilayer Networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[104] B. Klimt and Y. Yang. Introducing the Enron corpus. *CEAS*, 2004.

[105] J.M. Klusowski and Y. Wu. Estimating the number of connected components in a graph via subgraph sampling. *Accessed at https://arxiv.org/pdf/1801.04339.pdf on February 16, 2018*, 2018.

[106] E.D. Kolaczyk. *Statistical Analysis of Network Data*. Springer Series in Statistics. Springer, New York, 2009. Methods and models.

[107] E.D. Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis (Re)Visiting the Foundations*. SemStat Elements. Cambridge, 2017.

[108] E.D. Kolaczyk and G. Csárdi. *Statistical Analysis of Network Data with R*. Use R! Springer, 2014.

[109] P.N. Krivitsky and M.S. Handcock. A Separable Model for Dynamic Networks. *Journal of the Royal Statistical Society Series B*, 76(1):29–46, 2014.

[110] P.N. Krivitsky and E.D. Kolaczyk. On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical Science*, 30(2):184–198, 2014.

[111] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65, 2000.

[112] S. H. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.

[113] L. Li, D. Alderson, J.C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.*, 2(4):431–523, 2005.

[114] X. Li and K. Rohe. Central limit theorems for network driven sampling. *Accessed at arXiv:1509.04704*, 2015.

[115] L. Lovász. *Large Networks and Graph Limits*, volume 60 of *AMS Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.

[116] L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96:933–957, 2006.

[117] S. Mankad and G. Michailidis. Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world? *Annals of Applied Statistics*, 9(4):1950–1972, 2015.

[118] R. Martin and C. Liu. *Inferential Models: Reasoning with Uncertainty*. Chapman & Hall, 2016.

[119] N. Masuda and P. Holme (eds.). *Temporal Network Epidemiology*. Springer Nature, Singapore, 2017.

[120] P. McCullagh. What is a statistical model? *Ann. Statist.*, 30(5):1225–1310, 2002. With comments and a rejoinder by the author.

[121] J.L. Moreno. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Beacon House, 1934.

[122] S. Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System. *Accessed at https://bitcoin.org/bitcoin.pdf on February 13, 2018.*

[123] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256 (electronic), 2003.

[124] Y.C. Ng and R. Silva. A Dynamic Edge Exchangeable Model for Sparse Temporal Networks. *Accessed at arXiv:1710:04008*, 2017.

[125] K. Okike, K.T. Hug, M.S. Kocher, and S.S. Leopold. Single-blind vs Double-blind Peer Review in the Setting of Author Prestige. *Journal of the American Medical Association*, 316(12):1315, 2016.

[126] P. Orbanz. Comment on F. Caron and E.B. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society, Series B*, 79(5), 2017.

[127] P. Orbanz. Subsampling large graphs and invariance in networks. *arXiv:1710.04217*, 2017.

[128] P. Orbanz and D.M. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461, 2015.

[129] K. Palla, F. Caron, and Y.W. Teh. Bayesian nonparametrics for sparse dynamic networks. *arXiv:1607.01624*, 2016.

[130] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probab. Th. Relat. Fields*, 92:21–39, 1992.

[131] P.O. Perry and P.J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society, Series B*, 75:821–849, 2013.

[132] J. Pitman. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[133] S.I. Resnick. *Heavy Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, 2007.

[134] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[135] M. Sageman. *Understanding Terror Networks*. University of Pennsylvania Press, 2004.

[136] M. Schweinberger, P.N. Krivitsky, and C.T. Butts. Foundations of Finite-, Super-, and Infinite-Population Random Graph Inference. *Accessed at arXiv:1707.04800*, 2017.

[137] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[138] C.R. Shalizi and A. Rinaldo. Consistency under subsampling of exponential random graph models. *Annals of Statistics*, 41:508–535, 2013.

[139] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[140] T.A.B. Snijders. Stochastic Actor-Oriented Models for Network Dynamics. *Annual Review of Statistics and Its Application*, 4:343–363, 2017.

[141] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic block models for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[142] D.L. Sussman and E.M. Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *Accessed at arXiv:1702.03578*, 2017.

[143] S.K. Thompson and O. Frank. Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26:87–98.

[144] S.K. Thompson and G.A.F. Seber. *Adaptive Sampling*. Wiley, New York, 1996.

[145] A. Todeschini, X. Miscouridou, and F. Caron. Exchangeable Random Measures for Sparse and Modular Graphs with Overlapping Communities. *Accessed at arXiv:1602.0211*, 2016.

[146] R. van der Hofstad. *Random walks and complex networks*. Lecture notes. 2012.

[147] V. Veitch and D. Roy. The Class of Random Graphs Arising from Exchangeable Random Measures. *Accessed at arXiv:1512.03099*, 2015.

[148] V. Veitch and D.M. Roy. Sampling and Estimation for (Sparse) Exchangeable Graphs. *Accessed at arXiv:1611.00843*, 2016.

[149] P. Wan, T. Wang, R.A. Davis, and S.I. Resnick. Fitting the linear preferential attachment model. *Electronic Journal of Statistics*, 11(2):3738–3780, 2017.

[150] S.S. Wasserman and P.E. Pattison. Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika*, 61(3):401–425, 1996.

[151] D. Watts. *Six Degrees: The Science of a Connected Age*. W.W. Norton, 2004.

[152] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[153] S.A. Williamson. Nonparametric Network Models for Link Prediction. *Journal of Machine Learning Research*, 17:1–21, 2016.

[154] W. Willinger, D. Alderson, and J.C. Doyle. Mathematics and the Internet: A source of enormous confusion and great potential. *Notices Amer. Math. Soc.*, 56(5):586–599, 2009.

[155] P.J. Wolfe and S.C. Olhede. Nonparametric graphon estimation. *Available at arXiv:1309.5936*, 2014.

[156] E.P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 4(2):535–566, 2010.

[157] M. Xu, V. Jog, and P.-L. Loh. Optimal rates for community estimation on the weighted stochastic block model. *Accessed at http://www-stat.wharton.upenn.edu/˜ minx/docs/wsbm.pdf on February 22, 2018*, 2018.

[158] J. Yang, C. Han, and E. Airoldi. Nonparametric estimation and testing of exchangeable graph models. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, PMLR*, 33:1060–1067, 2014.

[159] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Detecting communities and their evolutions in dynamic social networks–a Bayesian approach. *Machine Learning*, 82(2):157–189, 2011.

[160] S. Young and E. Scheinerman. Random dot product graph models for social networks. *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, pages 138–149, 2007.

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

[162] Y. Zhang, E.D. Kolaczyk, and B.D. Spencer. Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. *Annals of Applied Statistics*, 9(1):166–199, 2015.