18

PROBABILISTIC NETWORK ANALYSIS

PHILIPPA PATTISON AND GARRY ROBINS

INTRODUCTION

The aim of this chapter is to describe the foundations of probabilistic network theory. We review the development of the field from an early reliance on simple random graph models to the construction of progressively more realistic models for human social networks. Hence, we show how developments in probabilistic network models are increasingly able to inform our understanding of the emergence and structure of social networks in a wide variety of settings.

SOCIAL NETWORKS

Growing numbers of social scientists from an increasingly diverse set of disciplines are turning their attention to the study of social networks. The precise reasons vary, but almost certainly, there are at least two key factors at work. The first is an increasing recognition that networks matter in many realms of social, political, and economic life. Networks both potentiate and constrain the social interactions that, for instance, underpin the dissemination of knowledge, the exercise of power and influence, and the transmission of communicable diseases. Ignoring the structured nature of these interactions often leads to erroneous conclusions about their consequences, as social scientists from a number of disciplines have repeatedly pointed out (e.g., Bearman, Moody, & Stovel, 2004; Kretzschmar & Morris, 1996). The second reason for a heightened focus on social networks is our increasing capacity to measure, monitor, and model social networks and their evolution through time and hence to draw social networks into a more general program for a quantitative social science. Probabilistic models for social networks have played—and will likely increasingly play—a vital role in these developments. In this chapter, we review the progress in attempts to develop probabilistic network models and point to areas of ongoing development.

WHAT IS DISTINCTIVE ABOUT MODELS FOR SOCIAL NETWORKS?

At the outset, it is important to recognize that social networks pose particular challenges as far as probability modeling is concerned. Unlike observations on a set of distinct actors, where an assumption of independent observations may often seem reasonable, social relationships are much less plausibly regarded as independent. Relational observations may share one or more actors and hence be subject to influences such as the goals and constraints of a particular actor.

Alternatively, they may be linked by other relationships (e.g., the relationship between actors i and j may be linked with the relationship between actors k and l by a relationship involving actors j and k) and hence dependent, for example, by virtue of competition or cooperation regarding relational resources involving actors j and k. As we see below, the development of probabilistic network models began with simple models that assumed independent relational ties, but empirical researchers quickly confronted the problem that social networks appeared to deviate from simple random structures in seemingly systematic ways. Hence, the story of the development of probabilistic network models is a story of alternatively probing and parameterizing progressively more complex systematicities in network structure.

Notation and Some Basic Properties of Graphs and Directed Graphs

We begin with some notations and some important definitions, referring the reader to Wasserman and Faust (1994); see also Bollobás (1998) for a fuller exposition of key concepts.

Graphs and Directed Graphs

We let $N = 1,2,\ldots,n$ be a set of network *nodes*, with each node representing a social *actor*. The actors are often persons but may also be groups, organizations, or other social entities. An observed social network may be represented as a *graph* $G = (N,E)$ comprising the *node set* N and the *edge set* E comprising all pairs (i,j) of distinct actors who are linked by a network tie. The tie, or edge, (i,j) is said to be *incident* with nodes i and j. In this case, the network ties are taken to be *nondirected*, with no distinction between the tie from actor i to actor j and the tie from actor j to actor i; in other words, the edges (i,j) and (j,i) are regarded as indistinguishable. If it is desirable to distinguish these ties—as it often is—the network may be represented instead by a *directed graph* (N,E) on N: The node set is then also N, and the *arc set* E is the set of all *ordered* pairs (i,j) such that there is a tie *from* actor i to actor j. The convention of using the term *edge* in the case of a nonordered pair (i.e., a *nondirectional* tie) and *arc* in the case of an ordered pair (i.e., a *directional* tie) is widely adopted by graph theorists, although network researchers are inclined to use the term *tie* interchangeably in both cases. In many cases, by convention, ties of the form (i,i), known as *loops*, are excluded from consideration.

Graphs and directed graphs can be conveniently represented by a *graph drawing*. The elements of the node set N are represented by points in the drawing, and a nondirected line connects node i and node j if (i,j) is an edge in the edge set E. In the case of a directed graph, an arc represented by a directed arrow is drawn from node i to node j if (i,j) is in the arc set E. Figures 18.1 and 18.2 show examples of a graph and a directed graph, respectively.

Order, Size, and Density

The *order* and *size* of a graph are defined to be the number of nodes and edges, respectively; likewise, the order and size of a directed graph are the number of nodes and arcs, respectively. For example, the graph in Figure 18.1 has order 17 and size 34; the directed graph in Figure 18.2 has order 37 and size 169. The size of a graph of order n varies between 0 for an *empty* graph and $n(n-1)/2$ for the *complete* graph of order n (i.e., the graph in which every pair of nodes is linked by an edge). For a directed graph of order n, size varies between 0 and $n(n-1)$. The *density* of a graph or directed graph is a ratio of its actual size to the maximum possible size for n nodes, and it is in the range [0, 1]. The densities of the graph and directed graph of Figures 18.1 and 18.2 are 0.25 and 0.13, respectively.

Adjacency Matrix

Graphs and directed graphs can be represented by a binary *adjacency matrix*. For example, in the case of a graph, we can define $\mathbf{x}$ to be an $n \times n$ matrix with entries $x_{ij} = 1$ if there is an edge between i and j and $x_{ij} = 0$ otherwise. Since $x_{ij} = 1$ if and only if $x_{ji} = 1$, $\mathbf{x}$ is necessarily a *symmetric* matrix. In the case of a directed graph, unit entries in $\mathbf{x}$ correspond to arcs in E (i.e., $x_{ij} = 1$ if and only if

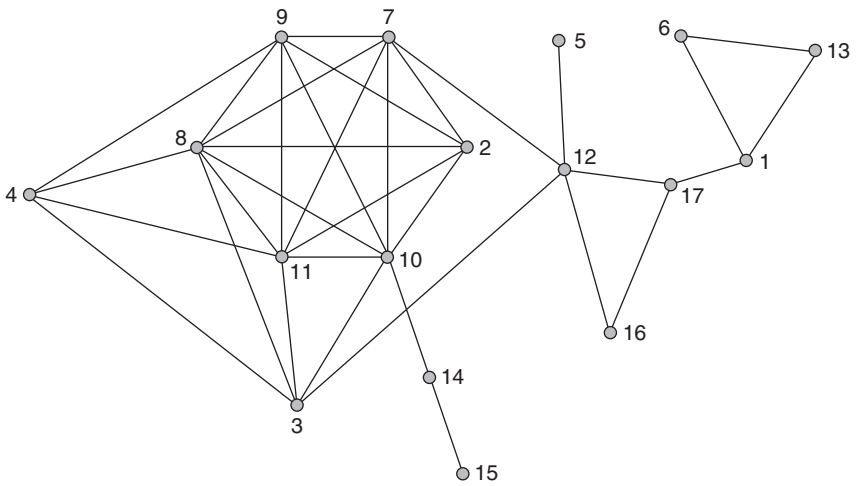

Figure 18.1 A graph on 17 nodes (mutual friendship network).

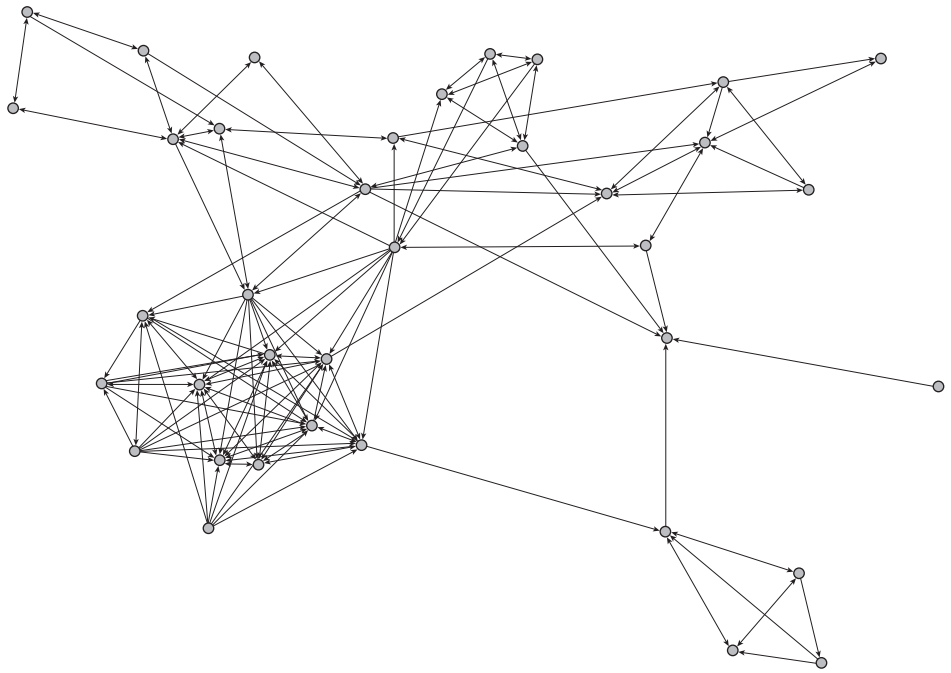

Figure 18.2 A directed graph on 37 nodes (reported collaboration network).

there is an arc from i to j, and $x_{ij} = 0$ otherwise). In this case, symmetry is not necessarily implied. The adjacency matrix corresponding to the graph of Figure 18.1 is shown in Table 18.1. Note that the size of a graph is $x_{++}/2$, whereas the size of a directed graph is x_{++}, where $x_{++} = \sum_{ij} x_{ij}$ is the sum of entries in the adjacency matrix.

Degree, Degree Sequence, and Degree Distribution

The number $k(i)$ of edges incident with a given node i is termed its *degree*: $k(i) = \sum_j x_{ij}$ is the sum of row i in the adjacency matrix $\mathbf{x}$. The *degree sequence* $(k(1), k(2), \ldots, k(n))$ of a

Table 18.1 Adjacency Matrix for Graph of Figure 18.1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	1	0	1	1	1	0	0	0	0	0
4	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
7	0	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
8	0	1	1	1	0	0	1	0	1	1	1	0	0	0	0	0	0
9	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0
10	0	1	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0
11	0	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0
12	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	1
13	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
17	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0

graph is the sequence of the degrees of its nodes, indexed by the labels $1,2,\ldots,n$ of nodes in N. The *degree distribution* is $(d_0,d_1,\ldots,d_{n-1})$, where d_k is the number of nodes in G of degree k. For example, the degree distribution of the graph in Figure 18.1 is shown in Figure 18.3. In a directed graph, the concept of degree is more complex, since there may be an arc directed from node j toward a given node i or away from node i toward node j, or there may be arcs in both directions between nodes i and j. We therefore characterize each node i in a directed graph by its outdegree $\text{out}_i = x_{i+} = \sum_j x_{ij}$, indegree $\text{in}_i = x_{+i} = \sum_j x_{ji}$, and mutual degree $\text{mut}_i = \sum_j x_{ij}x_{ji}$.

Subgraphs and the Dyad and Triad Census

Each subset S of the node set N of a graph $G = (N,E)$ gives rise to an *induced subgraph* H of G with node set S and edge set E' containing all edges in G that link pairs of nodes in S. More generally, any graph $H = (N',E')$ is a *subgraph* of G if $N' \subseteq N$ and $E' \subseteq E$. For example, the subgraph induced by the node set 1, 6, 13 of the graph of Figure 18.1 has edge set (1, 6),(1, 13),(6, 13); the graph comprising the node set 1, 6, 13 and the edge set (1, 6),(1, 13) is a subgraph of the graph of Figure 18.1 but not an induced subgraph. If every pair of nodes in the subgraph H is connected by an edge, then H is said to be a *clique*; for example, the subgraph induced by 2, 7, 8, 9, 10, 11 in Figure 18.1 is a clique of order 6.

A useful set of descriptive statistics for a graph or directed graph is a summary of the form of all of its small subgraphs. For example, the *dyad census* is a count of the number of each possible type of two-node induced subgraphs, and the *triad census* is the set of counts of three-node induced subgraphs. For example, the graph of Figure 18.1 has 102 null dyads and 34 linked dyads; its triad census comprises 279, 322, 49, and 30 induced three-node subgraphs with zero, one, two, and three edges, respectively. The dyad census and the triad census for directed graphs are defined similarly, but the number of forms of two-node and three-node subgraphs is greater in the directed graph case.

Implicit in the description of the dyad and triad census is the notion that two graphs (or subgraphs) can have the same form. We can make this notion more explicit by defining an isomorphic mapping between graphs. Specifically, two graphs $G = (N,E)$ and $H = (N',E')$ are *isomorphic* if there is a one-to-one mapping φ from N onto N' such that (i,j) is an edge in E if and only if $(\varphi(i),\varphi(j))$ is an edge in E'.

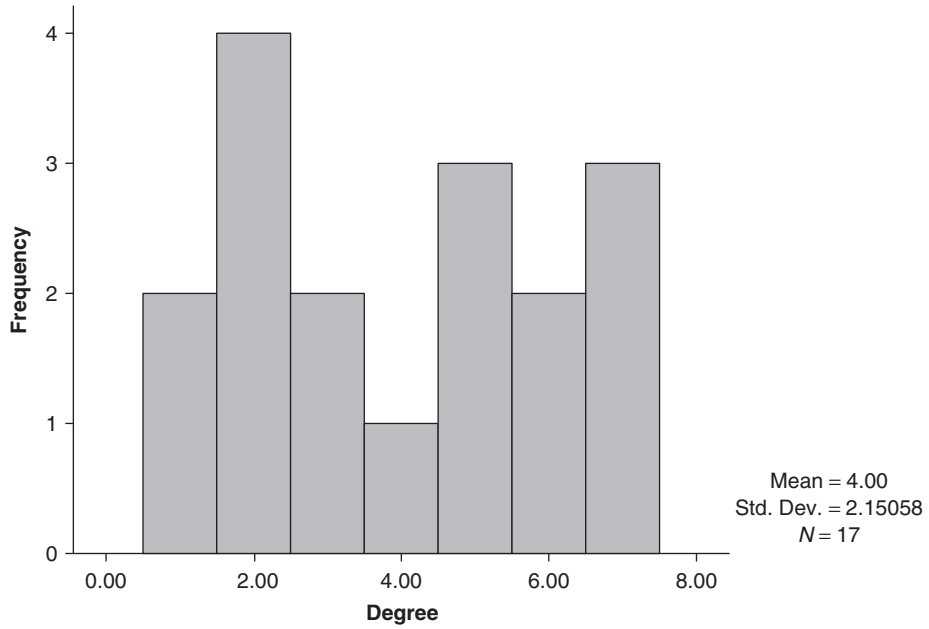

Figure 18.3 The degree distribution for the mutual friendship graph.

Paths, Reachability, and Connectedness

Social networks are often important to understand because social processes—such as the diffusion of information, the exercise of influence, and the spread of disease—are potentiated by network ties. Not surprisingly, therefore, pathlike structures in networks that might be associated with the flow of social processes are important concepts. A *path* from node i to node j is an ordered sequence $i = i_0, i_1, \ldots, i_l = j$ of distinct nodes in which each adjacent pair (i_{j-1}, i_j) is linked by an edge or an arc. The *length* of the path is l. If there is a path from a node i to a node j, then j is said to be *reachable* from i. If node j is the same as node i, then the path is termed a *cycle* of length l.

A *geodesic* from node i to another node j is a path of minimum length, and the *geodesic distance* d_{ij} from node i to node j is the length of the geodesic. If there is no path from i to j, the geodesic distance is *infinite*. The geodesic distance d_{ij} for distinct nodes i and j is either an integer in the range from 1 to $n-1$ or infinite. For a graph, geodesic distances are symmetric, that is $d_{ji} = d_{ij}$; this is not necessarily the case however, for directed graphs. The *geodesic distribution* of a graph or directed graph is the distribution of frequencies of geodesic distances—that is, the distribution of counts of the number of ordered pairs of nodes having each possible geodesic distance. The geodesic distribution of the graph of Figure 18.1 is presented in Figure 18.4 in the form of a histogram. The geodesic distribution can be seen as a useful summary of internode distances. Later, we refer to the quartiles of this distribution as simple summary statistics for internode distances.

If each node in a graph G is reachable from each other node, then G is *connected*. A *component* of G is a maximal connected subgraph—that is, a connected subgraph with vertex set W for which no larger set Z containing W is connected. The graph of Figure 18.1 is clearly connected.

In the case of a directed graph, we may also define a *semipath* from node i to node j as an ordered sequence $i = i_0, i_1, \ldots, i_l = j$ of distinct nodes in which either (i_{j-1}, i_j) or (i_j, i_{j-1}) is an arc. The *length* of the semipath is m. If each node in a directed graph G is reachable from each other node, then G is *strongly connected*. If there is a semipath from each node in G to each other node, then G is said to be *weakly connected*.

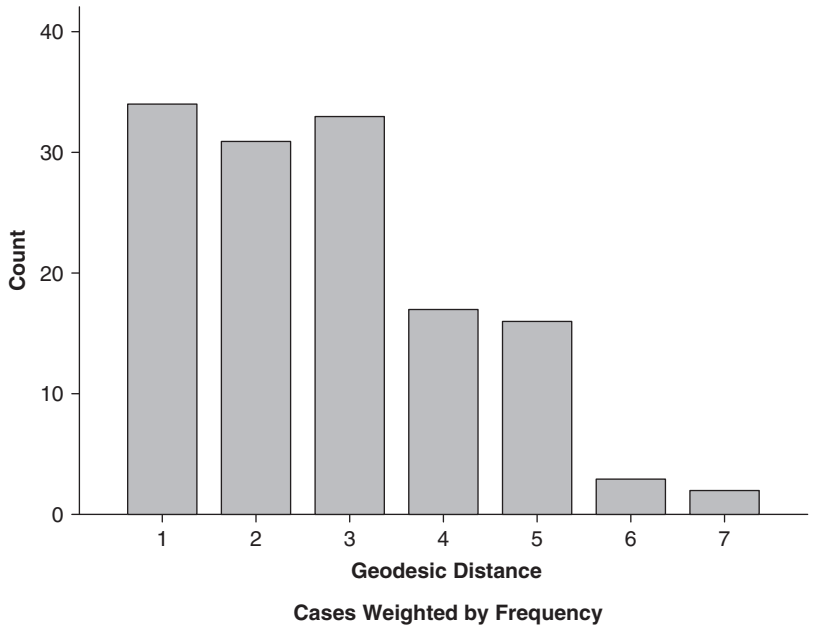

Figure 18.4 Geodesic distribution for the mutual friendship graph.

Simple Random Graphs and Directed Graphs

The Hungarian mathematicians Paul Erdős and Alfréd Rényi initiated an important approach to the study of random graph structures with a foundational series of papers beginning in 1959 (Erdős & Rényi, 1959). They introduced two primary random graph distributions on a fixed node set $N = 1, 2, \ldots, n$. The probability distributions are defined on the set of all graphs on n distinct nodes; this set contains $2^{n(n-1)/2}$ graphs, since each of the $n(n-1)/2$ pairs of nodes may or may not be linked by an edge. Each of the two random graph distributions that we introduce below associates a probability with every graph in this set.

$\mathcal{G}(n,p)$

In the first case, the edges of a graph are regarded as a set of independent Bernoulli variables. If we let X_{ij} denote the edge variable for the pair of nodes i and j and p be the (uniform) probability that the edge between i and j is present, then we can write $\Pr(X_{ij} = 1) = p$ and, hence, $\Pr(X_{ij} = 0) = 1 - p$. Since the edge variables are independent, it is then easy to write down the probability of any particular graph H of order n and size m:

$$\Pr(G = H) = p^m (1-p)^{m^* - m},$$

where $m^* = n(n-1)/2$ is the maximum number of edges in a graph of order n. The set of all possible graphs of order n and their corresponding probabilities is the *random graph distribution* $\mathcal{G}(n,p)$.

In the special case where $p = 0.5$, the probability of each graph H on n nodes is

$$\Pr(G = H) = (0.5)^m (0.5)^{m^* - m} = (0.5)^{m^*},$$

and hence, every graph on n nodes is equiprobable. This distribution is often termed the *uniform* random graph distribution and is denoted by U.

$\mathcal{G}(n,m)$

The second random graph distribution associates nonzero probabilities only with graphs of order n and size m; furthermore, every such graph is assumed to be equiprobable. Since there are

$n!/(m!(n-m)!)$ distinct graphs in the class, the probability of any particular graph of order n and size m is

$$\Pr(G = H) = m!(n-m)!/n!.$$

The distribution $\mathcal{G}(n,m)$ may be regarded as the uniform random graph distribution on n nodes, conditional on the property of having m edges. It may also be designated $U \mid x_{++} = m$. More generally, we can define a conditional uniform random graph distribution in terms of any graph property $\mathcal{Q}$ (Bollobás, 1985). If $\mathcal{Q}$ is a subset of all possible graphs on n nodes, then the distribution $U \mid \mathcal{Q}$ assigns equal probabilities (viz., $1/\mid \mathcal{Q} \mid$) to all graphs in the subset $\mathcal{Q}$ and zero probability to all graphs not in $\mathcal{Q}$. We term $U \mid \mathcal{Q}$ the uniform random graph distribution *conditional on* $\mathcal{Q}$.

Some Results for Simple Random Graphs

The field of random graphs in $\mathcal{G}(n,p)$ and $\mathcal{G}(n,m)$ has grown rapidly since its inception. Much work has explored the features of various random graph classes, particularly as n becomes large, and the way in which these features change, often very rapidly, as a function of p. Bollobás (1998) contains an excellent introduction to the field, as does the review by Albert and Barabási (2002); here, we illustrate just two aspects of this literature by considering the expected values of some statistics in $\mathcal{G}(n,p)$ and by reviewing properties of a graph as a function of p in $\mathcal{G}(n,p)$.

We begin by determining the expected number of cliques in a graph. Let $Y_s = Y_s(G)$ be the number of cliques of order s in the graph G. Then, the expected value of Y_s can readily be computed as

$$E(Y_s) = (n!/[s!(n-s)!])p^u,$$

where $u = s(s-1)/2$ is the number of edges in a clique of order s (e.g., see Bollobás, 1998). In $\mathcal{G}(17,0.25)$, for example, the expected number of cliques of order 3 is 10.6, and the expected number of cliques of order 4 is 0.58. The graph of Figure 18.1 has 17 nodes and a density of 0.25, and it is therefore interesting to compare the expected values for $\mathcal{G}(17,0.25)$ with those observed for the graph of Figure 18.1—namely, 34 cliques of order 3 and 18 cliques of order 4.

Indeed, if F is any subgraph with s nodes and t edges and Y_F is the number of subgraphs of G that are isomorphic to F, then the expected value of Y_F can readily be shown to be

$$E(Y_F) = (n!/[(n-s)!a])p^t,$$

where a is the number of distinct ways in which the nodes of the graph F can be labeled with the integers $1,2,\ldots,s$ to yield the same graph. For example, in the case of a cycle of order s, F has s edges and s nodes, and there are $2s$ distinct ways in which nodes can be labeled to yield the same graph; hence,

$$E(Y_F) = (n!/[(n-s)!2s])p^s.$$

The expected number of *induced* subgraphs isomorphic to F may be similarly derived:

$$E(Y_F) = (n!/[(n-s)!a])p^t(1-p)^{s(s-1)/2-t}.$$

The latter formula may be used, for example, to compute the expected triad census for a graph of a given order and density.

The expression for the expected number of subgraphs isomorphic to F in $\mathcal{G}(n,p)$ allows us to explore features of random graphs as n tends to infinity. Since

$$E(Y_F) = (n!/[(n-s)!a])p^t \approx n^s p^t/a,$$

it is clear that if $p = cn^{-s/t}$, then $E(Y_F) \approx c^t/a$ and the expected number of subgraphs isomorphic to F, denoted by $\lambda = c^t/a$, is a finite number. If, however, $pn^{s/t}$ tends to 0 or ∞ as n tends to ∞, then the probability that a random graph in $\mathcal{G}(n,p)$ contains at least one subgraph F converges to 0 or 1, respectively (e.g., Albert & Barabási, 2002). Thus, $p = cn^{-s/t}$ is a critical probability below which graphs with large n rarely contain F and above which they almost certainly do. An important set of results in random graph theory documents the many graph properties that show this form of rapid transition from being very unlikely to very likely as a function of the edge probability p.

Application of this approach allows us to infer for large n some expected features of random graphs in $\mathcal{G}(n,p)$ as a function of p, relative to n. Thus, for example, if $p < 1/n$, then almost every

graph in $\mathcal{G}(n,p)$ comprises a number of components, each without any cycles; if p lies between $1/n$ and $(\ln n)/n$, then almost every graph has a so-called giant component (i.e., a component including a large proportion of the nodes in N); and if $p > (\ln n)/n$, then almost every graph is connected (e.g., see Albert & Barabási, 2002).

Random directed graph distributions may be similarly defined, though interest in them has largely come from social scientists with applications to network data in mind. It is to this literature that we now turn.

Applications of Random Graph and Directed Graph Distributions to Social Network Data

Before describing the application of random graphs to social networks, it is important to say something about typical sources of social network data (e.g., Wasserman & Faust, 1994). A common method of measuring social networks is to survey all members of a circumscribed population about their ties. In this case, ties are typically directional, and there may or may not be a limit imposed on the maximum number of ties reported by each respondent. Occasionally in this case, it is fruitful to consider the graph constructed from mutual ties only. For example, the graph of Figure 18.1 is the set of mutual ties observed among the girls in a Grade 8/9 high school class, obtained in response to the question "Who are your best friends in the class?" Networks are also commonly inferred from archival data, such as communication logs or membership or attendance lists. Less common strategies include direct observation and more elaborate survey techniques.

Application of Directed Random Graph Distributions

In the 1930s, the psychiatrist Jacob Moreno and colleagues reported using random directed graph distributions to compute quantities such as the expected number of mutual ties (i.e., $\sum_{i,j} X_{ij}X_{ji}/2$), where X_{ij} denotes the random variable for the tie from node i to node j. Moreno and colleagues also calculated properties of the indegree distribution in a random directed graph distribution (see, e.g., the very interesting historical account in Freeman, 2004).

Consider, for example, the random graph distribution $\mathcal{DG}(n,p)$ with a fixed set of n nodes and uniform but unknown arc probability p; this is the directed graph analog of $\mathcal{G}(n,p)$. The expected number of mutual ties for directed graphs in this distribution is $n(n-1)p^2/2$, and the expected number of nodes with indegree k is $np^k(1-p)^{n-1-k}$. If a social network with n nodes and $x_{++} = \sum_{ij} x_{ij}$ arcs has been observed, then the arc probability p can be estimated from the network data as $p^* = x_{++}/(n[n-1])$. This estimate of the arc probability can be used to compute the expected number of mutual ties and the expected number of nodes with each possible indegree k, on the assumption that the observed network was generated from $\mathcal{DG}(n,p^*)$. These expected values can be compared with the observed number of mutual ties and the observed indegree distribution in the network $\mathbf{x}$. If the observed values are markedly different from the expected ones, then it can be argued that there is reason to question the suitability of the assumption of independent random arcs with uniform probability that underpinned the computation of expected values.

The computations by Moreno and colleagues revealed what would later become a very common finding: that the observed number of mutual ties in an observed human social network is much greater than the number expected on the basis of arc probability alone and the indegree distribution is more heterogeneous than expected—that is, there are more nodes with very low and very high indegree than expected. These and similar findings suggest that tendencies toward mutuality and heterogeneity in partner "attractiveness" are systematic features of observed social networks comprising ties of affiliation.

Most important for our purposes here, Moreno introduced the idea of using random directed graph distributions as *null distributions*. The features of this null distribution could be compared with features of an observed network, a comparison enabling researchers to identify the ways in which the observed network appeared to be systematically different. In this early example, and in many to follow, it was important that the expected features of the null distribution could be

derived mathematically. Much later, when fast computers and more versatile simulation algorithms were introduced, this restriction could be relaxed, but it was an important reason for the focus of early applications on this "null distribution" approach. Moreover, although this early application assumed very simple null distributions (such as independent arcs with uniform tie probability), more complex distributions were soon developed. Indeed, the strategy continues to be used in new ways (e.g., Bearman et al., 2004; Pattison, Wasserman, Robins, & Kanfer, 2000) and to be rediscovered in new fields (e.g., Milo et al., 2002). It remains an important means by which some of the systematic structural properties of human social networks can be and have been uncovered.

Holland, Leinhardt, and colleagues were responsible for developing a number of important elaborations of this basic strategy. For example, Holland and Leinhardt (1975) computed the expected mean vector and variance-covariance matrix for the triad census in the uniform random directed graph distribution $U \mid$ mut, asym, null[1] conditional on fixed numbers mut, asym, and null of mutual, asymmetric, and null ties, respectively (mut = $\sum_{ij} x_{ij}x_{ji}/2$, asym = $\sum_{ij}[x_{ij}(1 - x_{ji}) + x_{ji}(1 - x_{ij})]$, null = $\sum_{ij}(1 - x_{ij})(1 - x_{ji})$). In other words, they computed the expected distribution of the triad census while conditioning on the dyad census. This allowed them to construct a test statistic for any linear combination of triad counts and, hence, assess whether the observed combination of triad counts is in the upper or lower tail of the expected distribution. For example, they could test for the presence of *transitivity* (i.e., the property that arcs from nodes i to j and from j to k are accompanied by an arc from i to k).

Similar calculations can be made for other distributions of possible interest, including the uniform distribution $U \mid \{x_{i+}\}$,mut conditional on the outdegrees of each node in the directed graph as well as the number of mutual ties. Of course, for some desirable combinations, such as $U \mid \{x_{i+}\}, \{x_{+i}\}$,mut, the calculations are very difficult and have prompted alternative parametric approaches. In some of these difficult cases, clever simulation strategies have been devised to circumvent the difficult mathematics. For example, Snijders (1991) used an importance-sampling approach to simulate $U \mid \{x_{i+}\}, \{x_{+i}\}$, and McDonald, Smith, and Forster (2007) have described a Markov chain Monte Carlo algorithm to simulate the distribution $U \mid \{x_{i+}\}, \{x_{+i}\}$, mut.

[1]In fact, Holland and Leinhardt (1975) termed this the $U \mid$ MAN distribution, but we have attempted to keep notation consistent in the chapter.

Biased Nets

One other early probabilistic approach deserves mention. In a series of papers, Rapoport and colleagues developed the theory of *biased nets*—that is, random networks with biases toward symmetry, transitivity, and other features characteristic of observed social networks (Rapoport, 1957). Although a full and satisfactory mathematical treatment proved elusive, Rapoport and colleagues employed their conceptualization of biased nets to conduct some illuminating studies of the connectivity structure of a large friendship network (e.g., Rapoport & Horvath, 1961). In part, their work can be seen as the intellectual precursor to the more general probabilistic developments described below, but a different framing of the "biases" has proved more useful.

The p_1 Model

Comparison of observed social networks with random directed graph distributions consistently revealed a greater than expected number of mutual ties and greater than expected degree heterogeneity. As a consequence, it was felt desirable to compare observed networks with graph distributions that resembled observed networks in these fundamental respects. The problem of satisfactorily simulating the random graph distribution conditional on the number of mutual ties and the in-degree and out-degree sequences is arguably still not resolved. In the meantime, Holland and Leinhardt (1981) developed an alternative approach: a probability model that parameterized these tendencies. This model was an

important step toward the development of a more general framework.

The p_1 model developed by Holland and Leinhardt (1981) assumes *independent dyads* $D_{ij} = (X_{ij}, X_{ji})$. The distribution of the entire network $\mathbf{X} = [X_{ij}]$ can then be determined by specifying the probability of each possible dyadic form for D_{ij} since the probability of the entire network $\mathbf{X}$ is the product of the dyad probabilities. The individual dyad probabilities can be expressed in terms of the probability of occurrence of a mutual dyad, an asymmetric dyad, and a null dyad. Thus, we define

$$\begin{aligned}\Pr(D_{ij} = (1,1)) &= m_{ij} = m_{ji},\\ \Pr(D_{ij} = (1,0)) &= a_{ij}, \text{ and}\\ \Pr(D_{ij} = (0,0)) &= n_{ij} = n_{ji},\end{aligned}$$

where $m_{ij} + a_{ij} + a_{ji} + n_{ij} = 1$ for all $i \neq j$.

The resulting probability distribution

$$\Pr(\mathbf{X} = \mathbf{x}) = \prod_{i<j} m_{ij}^{X_{ij}X_{ji}} \prod_{i\neq j} a_{ij}^{X_{ij}(1-X_{ji})} \prod_{i<j} n_{ij}^{(1-X_{ij})(1-X_{ji})}$$

may then be reexpressed in the exponential form

$$\Pr(\mathbf{X} = \mathbf{x}) = K \exp[\sum_{i<j} \rho_{ij} X_{ij} X_{ji} + \sum_{ij} \theta_{ij} X_{ij}],$$

where, for all $i \neq j$,

- $\rho_{ij} = \log\{m_{ij} n_{ij} / (a_{ij} a_{ji})\}$ is an index of reciprocity,
- $\theta_{ij} = \log\{a_{ij} / n_{ij}\}$ is a log-odds measure of the probability of an asymmetric dyad between i and j, and
- $K = \prod_{i<j}[1/(1 + \exp(\theta_{ij}) + \exp(\theta_{ji}) + \exp(\rho_{ij} + \theta_{ij} + \theta_{ji}))]$ is a normalizing quantity.

Holland and Leinhardt added two useful restrictions to this general dyad-independent model. The first was that the reciprocity parameter ρ_{ij} is a constant for all dyads; that is, $\rho_{ij} = \rho$ for all $i \neq j$. The second was that the parameter θ_{ij} depended additively on the propensity of arcs to emanate from node i and the propensity of arcs to have node j as a target; in other words,

$$\theta_{ij} = \theta + \alpha_i + \beta_j, \quad \text{for } i \neq j.$$

The resulting model is termed the p_1 model:

$$\begin{aligned}p_1(\mathbf{x}) &= \Pr(\mathbf{X} = \mathbf{x})\\ &= K \exp\left[\rho \sum_{i,j} X_{ij} X_{ji} + \theta X_{++} + \sum_i \alpha_i X_{i+} + \sum_i \beta_i X_{+i}\right].\end{aligned}$$

The parameters ρ and θ can be interpreted as uniform *reciprocity* and *density* parameters, and the node-dependent parameters α_i and β_i reflect the *expansiveness* and *attractiveness*, respectively, of each node i.

The development of the p_1 model was an important step in probabilistic network theory, not the least because much of the machinery of statistical modeling could be brought to bear on the problem of assessing model adequacy. The model could be estimated from data, and its goodness of fit could be subjected to careful scrutiny, as Brieger (1981) demonstrated. Such scrutiny led to the recognition that observed networks often exhibited structural properties not captured by the parameters of the p_1 model and spawned the development of two important lines of further model development.

Latent Variable Models

The first line of development is a series of latent variable models in which the assumption of independent dyads is replaced by an assumption of independent dyads conditional on unobserved variables representing some potential underlying structure.

One such model was inspired by the concept of structural equivalence in a graph (Lorrain & White, 1971). Two nodes are structurally equivalent if they have identical patterns of relationships to other nodes. The concept of structural equivalence has been very influential in the social networks literature because it can be used to represent the idea that two actors have the same social position in a network; that is, that they are indistinguishable from a relational point of view.

Formally, two nodes i and j are *structurally equivalent* in a directed graph G with adjacency matrix $\mathbf{x}$ if $x_{ik} = x_{jk}$ and $x_{ki} = x_{kj}$ for all nodes $k \neq i, j$ in N. Structurally equivalent nodes can

be partitioned into *blocks*. Nowicki and Snijders (2001) assumed that the blocks to which nodes belong are unobserved. They defined a set of independent and identically distributed latent random variables $\mathbf{Z} = [Z_i]$, where Z_i denotes the *block* of node i and $\Pr(Z_i = k) = \theta_k$. They assumed the dyads $D_{ij} = (X_{ij}, X_{ji})$ to be conditionally independent given the blocks and the probability that a dyad has a particular relational form to depend only on the (unobserved) blocks of the nodes. In other words,

$$\Pr(D_{ij} = \boldsymbol{a} \mid \mathbf{Z} = \mathbf{z}) = \eta_a(z_i, z_j),$$

where $\boldsymbol{a}$ is a vector of possible values for the dyad, with $\boldsymbol{a} \in \{(1,1),(1,0),(0,1),(0,0)\}$ for a directed graph and $\boldsymbol{a} \in \{(1,1),(0,0)\}$ for a graph, and $\eta_a(z_i, z_j)$ is the block-dependent probability of observing the vector $\boldsymbol{a}$.

In this model, two nodes i and j are *stochastically equivalent* if they belong to the same block and hence the same dyad probabilities ($\Pr(D_{ik} = \boldsymbol{a} \mid \mathbf{Z} = \mathbf{z}) = \Pr(D_{jk} = \boldsymbol{a} \mid \mathbf{Z} = \mathbf{z})$) for all nodes k. Since the dyads are assumed to be conditionally independent given the blocks $\mathbf{Z}$, the joint distribution of the D_{ij} given $\mathbf{Z}$ is the product of the conditional dyad probabilities. Nowicki and Snijders (2001) developed a Bayesian approach to the estimation of θ and η and, hence, the computation of the posterior probabilities that any pair of nodes are in the same block and that a dyad has any particular relational form.

Several other important latent variable models have been developed for particular types of social networks that are likely to reflect some form of proximity among actors, such as friendship or collaboration. In cases such as these, it may be reasonable to assume that tie probabilities are monotonically related to proximity in a latent space. For example, Hoff, Raftery, and Handcock (2002) proposed a model that assumes that nodes have latent locations in some low-dimensional Euclidean space and that given these latent locations, tie variables are conditionally independent. Schweinberger and Snijders (2003) developed a similar model based on an ultrametric rather than Euclidean space; in their model, every pair of nodes is associated with an unobserved distance in an ultrametric space corresponding to a discrete hierarchy of "settings," and tie probabilities are conditionally independent given these latent ultrametric distances. (Distances in an ultrametric space satisfy the *ultrametric inequality*; i.e., $d(i,j) \leqslant \max\{d(i,k), d(j,k)\}$ for any triple of nodes i, j, k.) They developed approaches for estimating the unobserved ultrametric distances. Handcock, Raftery, and Tantrum (2005) have recently extended Hoff et al.'s (2002) model by assuming that the latent locations are drawn from a finite mixture of multivariate normal distributions, each of which represents a different group of nodes.

Markov Random Graphs

The second recent line of development has been to build probabilistic network models in which conditional dependencies among tie variables are permitted. This work began with the recognition by Frank and Strauss (1986) that a general approach for modeling interactive systems of variables (Besag, 1974) could be usefully applied to the problem of modeling systems of interdependent network tie variables on a fixed set of nodes. This was an important step because it permitted models to go beyond the limiting assumption of dyad independence in quite a general way. Frank and Strauss (1986) introduced a *Markov dependence* assumption for network tie variables: Two network tie variables were assumed to be conditionally independent given the values of all other network tie variables, unless they had a node in common. Thus, whereas a tie between nodes i and j was assumed to be conditionally independent of ties involving all other distinct pairs of nodes k and l, it could be conditionally dependent on *any* other ties involving i and/or j.

Assumptions about which pairs of tie variables are conditionally dependent, given the values of all other tie variables, can be represented as a dependence graph. The node set of the *dependence graph* D is the set of tie variables $\{X_{ij}\}$, and two tie variables are joined by an edge in D if they are assumed to be conditionally dependent given the values of all other tie variables. In the case of $\mathcal{G}(n,p)$, D is an empty graph since all pairs of variables are assumed to be mutually independent. In the Markov case, the variable X_{ij} is connected to X_{ik} and X_{jk} for all $k \neq i$ or j, and the dependence graph is connected. Figure 18.5

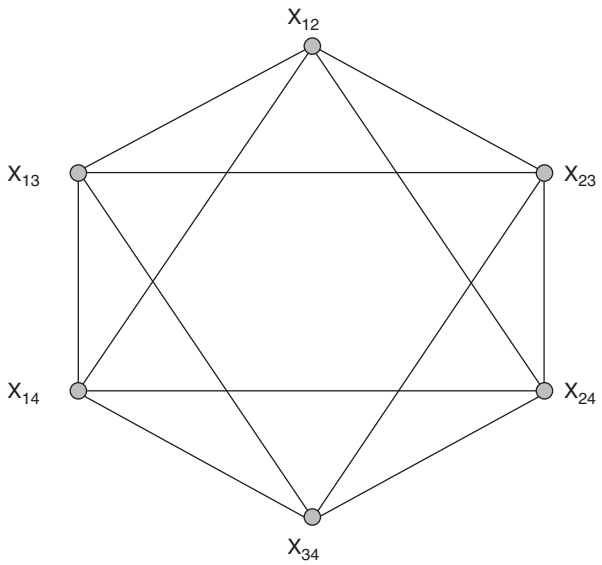

Figure 18.5 Dependence graph for a Markov random graph on four nodes.

shows the Markov dependence graph for a random graph of order 4.

As Frank and Strauss originally outlined, the consequences of any proposed assumptions about potential conditional dependencies among network tie variables can be inferred from the Hammersley-Clifford theorem (Besag, 1974). The theorem establishes a model for the interacting system of tie variables in terms of parameters that pertain to the presence or absence of certain configural forms in the network. The model, known as an *exponential random graph model*, takes the general form

$$\Pr(\mathbf{X} = \mathbf{x}) = \exp\left(\sum_A \gamma_A z_A(\mathbf{x})\right)/\kappa,$$

where A is a subset of tie variables (defining a potential network *configuration*), γ_A is a model parameter associated with the configuration A (to be estimated) and is nonzero only if the subset A is a clique in the dependence graph D, $z_A(\mathbf{x}) = \prod_{X_{ij} \in A} x_{ij}$ is the sufficient statistic corresponding to the parameter γ_A and indicates whether or not all tie variables in the configuration A have values of 1 in the network $\mathbf{x}$, and κ is a normalizing quantity.

To reduce the number of model parameters, Frank and Strauss (1986) introduced a *homogeneity* constraint that parameters for isomorphic configurations are equal. With this constraint, there is a single parameter $\gamma_{[A]}$ for each class $[A]$ of isomorphic configurations that correspond to cliques in the dependence graph. The sufficient statistic in the model corresponding to the class $[A]$ is then

$$Z_{[A]}(\mathbf{x}) = \sum_{A \in [A]} \prod_{X_{ij} \in A} x_{ij},$$

that is, a count of all observed configurations in the graph $\mathbf{x}$ that are isomorphic to the configuration corresponding to A. For example, in the case of a homogeneous Markov random graph, it is readily seen that cliques A in the dependence graph D correspond to graph configurations that are edges, stars, and triangles (see Figure 18.6), and the model therefore takes the form

$$\Pr(\mathbf{X} = \mathbf{x}) = \exp(\theta L(\mathbf{x}) + \sum_k \sigma_k S_k(x) + \tau T(\mathbf{x}))/\kappa,$$

where $L(\mathbf{x})$, $S_k(\mathbf{x})$, and $T(\mathbf{x})$ are the number of edges, k-stars $(2 \leqslant k \leqslant n-1)$, and triangles in the network $\mathbf{x}$ and θ, $\sigma_k (2 \leqslant k \leqslant n-1)$, and τ are the corresponding parameters.

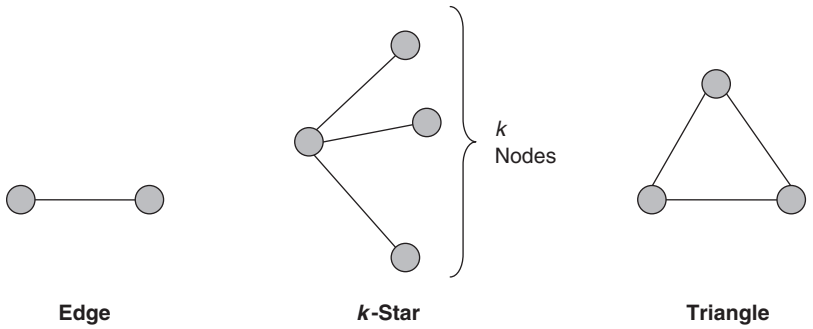

Figure 18.6 Markov model configurations: edges, stars, and triangles.

In many circumstances, the parameters may be interpreted by observing that if a configuration class $[A]$ has a large positive (or negative) parameter in the model, then the presence of many configurations in the class enhances (or reduces) the likelihood of the overall network, net of the effect of all other configurations. It should be noted, though, that the function relating the value of one of the model's parameters, say $\lambda_{[A]}$, to the expected value of the corresponding sufficient statistic $z_{[A]}(\mathbf{x})$ may be markedly nonlinear and exhibit a sharp and rapid transition from lower average counts to higher average counts, with a relatively small change in the parameter (holding constant the values of all other model parameters). For example, the expected number of triangles in a graph as a function of the triangle parameter τ is shown for a graph on 17 nodes in Figure 18.7. The values of the parameters θ, σ_2, and σ_3 are fixed at $-1.2558, -0.0451,$ and -0.1084, respectively, and τ takes values in the range $[0.05, 1.70]$. Figure 18.7 shows the distribution of the triangle statistic in the form of a box plot for each value of τ. It can be seen that for low values of τ, small increases in τ are associated with small and steady increases in the triangle statistic. As τ approaches 1.40, though, the impact of small changes in τ increases rapidly in magnitude, and there is a sharp transition to a higher value of the triangle statistic. Near the point of transition, the triangle statistic may take values typical of the graphs on either side of this apparent threshold. This form of nonlinear relationship is common, and the location of this threshold and the sharpness of the rise in the region of greatest sensitivity are likely to depend on other parameter values.

It is important to emphasize that even though this model is well understood in the case where only the parameter θ is nonzero (since this is just the model $\mathcal{G}(n,p)$ with $p = \exp(\theta)/[1+\exp(\theta)]$), more complex instantiations can be seen as models for self-organizing network processes (Robins, Pattison, & Woolcock, 2005). Robins et al. (2005) have demonstrated that specific sets of parameter values for the homogeneous Markov model can characterize very diverse network structures, including small worlds, caveman worlds, long-path worlds, and so on.

For some parameter values, the model may accord very high probability to a small set of graphs and very low probability to the rest, as Handcock (2004) and Snijders (2002) have demonstrated. Handcock termed these models *near-degenerate*. For detailed investigation of the behavior of specific models, see Handcock (2004), as well as Park and Newman (2004) and Burda, Jurkiewicz, and Krzywicki (2004).

Model Simulation

To understand properties such as near-degeneracy of the exponential random graph model $\Pr(\mathbf{X} = \mathbf{x}) = \exp(\sum_A \gamma_A z_A(\mathbf{x}))/\kappa$, it is helpful to be able to simulate it efficiently (i.e., to draw graphs $\mathbf{x}$ with probability $Pr(\mathbf{X} = \mathbf{x})$), and this generally means circumventing the need to compute the normalizing quantity κ, since κ is a function of *all* graphs in the distribution. As Strauss (1986) and others have observed, the *Metropolis algorithm* can be used for this purpose. The algorithm sets up a Markov chain on the space of all possible graphs of order n in such

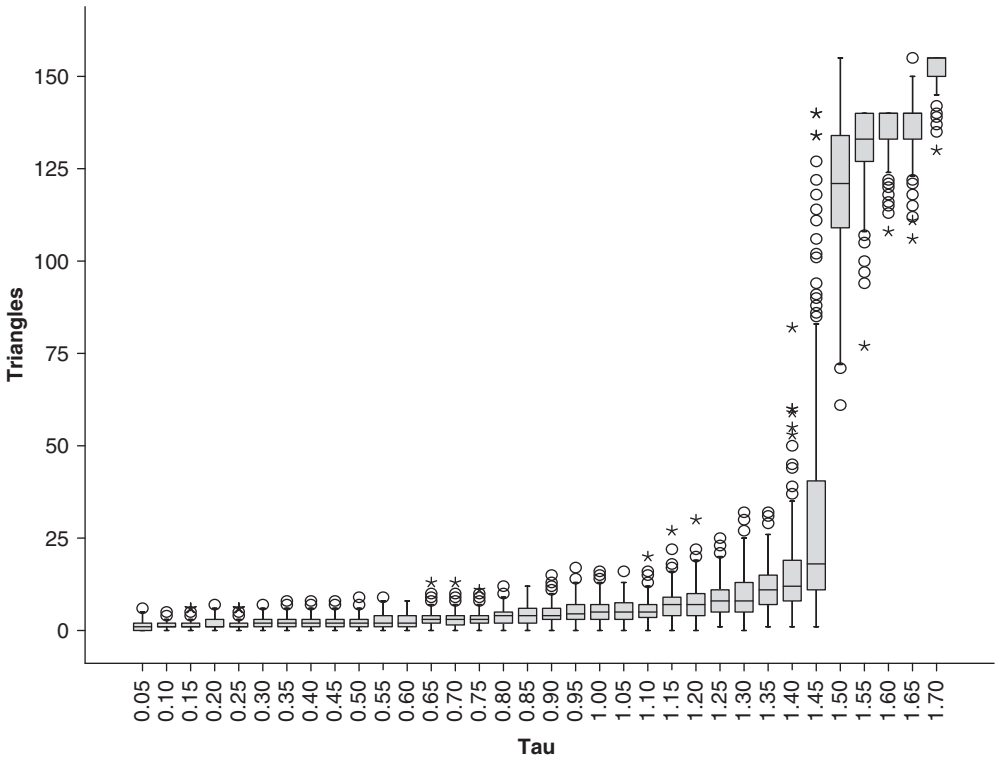

Figure 18.7 Boxplots for the triangle statistic as a function of the triangle parameter for a Markov model on a graph of 17 nodes ($\theta = -1.2558$, $\sigma_2 = -0.0451$, $\sigma_3 = -0.1084$, τ between 0.00 and 1.75).

a way that the Markov chain has the model as its stationary distribution. It may be described as follows:

1. Begin with some graph $\mathbf{x}$.
2. At each step, select an edge at random, say the (i, j) edge, and let $\mathbf{x}'$ be the graph that is identical to $\mathbf{x}$, except that the edge from i to j is switched to absent if it is present in $\mathbf{x}$ or to present if it is absent in $\mathbf{x}$.
3. Replace $\mathbf{x}$ by $\mathbf{x}'$ with probability $\min[1, \exp\{\sum_A \gamma_A(z_A(\mathbf{x}') - z_A(\mathbf{x})\}]$.
4. Return to Step 2, unless some specified target number of steps has been taken.

To sample from $\Pr(\mathbf{X} = \mathbf{x})$, it is usual to begin sampling graphs from the chain after some initial number of steps have been completed (the *burn-in* period); graphs are then sampled at a rate that may depend on n. Many variations of this approach may also be used; a valuable discussion may be found in Snijders (2002).

Estimation of Model Parameters

In many settings, primary interest lies in estimating exponential random graph model parameters from observed network data. For example, given the graph of Figure 18.1, there may be interest in estimating the parameters of a Markov model from which it might have been generated. In the early applications of these models to observed data, an approximate form of estimation known as pseudolikelihood estimation was often used (Strauss & Ikeda, 1990; Wasserman & Pattison, 1996), even though the properties of the estimates were not well understood. Initial attempts to apply the very promising approach of Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) were not always

successful because the properties of the models under consideration were not always fully appreciated, as Snijders (2002) and Handcock (2004) demonstrated. However, with a growing understanding of model properties and more careful attention to model adequacy, substantial progress has now been made in implementing MCMCMLE approaches (see Handcock, Hunter, Butts, Goodreau, & Morris, 2004; Snijders, 2002).

As an example of MCMCMLE, we estimate the parameters of the model

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{\exp(\theta L(\mathbf{x})+\sigma_2 S_2(\mathbf{x})+\sigma_3 S_3(\mathbf{x})+\tau T(\mathbf{x}))}{\kappa}$$

for the mutual friendship network of Figure 18.1 using the approach proposed by Snijders (2002). The resulting estimates and estimated standard errors for the parameters θ, σ_2, σ_3, and τ are shown in Table 18.2.[2] Also displayed in Table 18.2 are convergence t statistics, computed as the difference between the observed value of the sufficient statistic for a parameter and its average simulated value, divided by the standard deviation of simulated values. If the estimated value is indeed the maximum likelihood estimate, the simulated values should be centered on the observed value and the t statistics should all be small, preferably below 0.1 (Snijders, 2002). It can be seen from Table 18.2 that the t statistics satisfy this requirement. Although the edge, 2-star, and 3-star parameters are negative and within about 1 standard error of 0, the triangle parameter is positive and approximately seven times its estimated error. This suggests that, other graph features (edges, 2-stars, and 3-stars) being equal, graphs with more triangles are more likely. That such a model is needed for the graph of Figure 18.1 is consistent with the earlier computations based on $G(17, 0.25)$.

Goodness of Fit

A good statistical model should not be unnecessarily complex, but it should be adequate: that is, the data should resemble realizations from the model in many important respects. We can assess model adequacy by comparing the observed network with graphs generated by the model in features that are not necessarily parameterized within the model. What is important in such comparisons is very much a function of the modeling context, but there are often good reasons to require that the model captures the degree of clustering in a network, the distribution of degrees, and the connectivity structure that is represented by the geodesic distribution (e.g., Goodreau, 2007; Robins, Snijders, Wang, Handcock, & Pattison, 2007). It is important to note that only some of these characteristics need be associated with model parameters; others might be seen as consequences of these parameterized tendencies.

For example, if we simulate the model

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{\exp(\theta L(\mathbf{x})+\sigma_2 S_2(\mathbf{x})+\sigma_3 S_3(\mathbf{x})+\tau T(\mathbf{x}))}{\kappa}$$

using the parameter estimates in Table 18.2, we can not only compare the observed graph with the simulated graph in terms of its sufficient statistics (viz., the number of edges, 2-stars, 3-stars, and triangles), but we can also make the comparison in relation to any unmodeled network characteristic, such as the number of nodes of degree 4 or more, the number of geodesic distances of length 3, and so on.

Table 18.3 summarizes these comparisons for the graph of Figure 18.1 and the parameter estimates of Table 18.2. It can be seen that the t statistics are all less than 1 for

- the *local clustering coefficient* (the average across all nodes i of the proportion of pairs of nodes j and k incident with $i(x_{ij}=1=x_{ik})$ that are themselves connected $(x_{jk}=1)$,
- the *global clustering coefficient* (the proportion of the triples of nodes $\{i,j,k\}$ with $x_{ij}=1=x_{ik}$ for which $x_{jk}=1$),
- the *standard deviation* of the degree distribution, and
- the *skewness coefficient* of the degree distribution.

[2]The estimation was conducted using PNet (Wang, Robins, & Pattison, 2006), an implementation of the estimation approach in Snijders (2002). Retrieved from http://www.sna.unimelb.edu.au/pnetpnet.html.

Table 18.2 MCMCMLEs for Markov Model of the Graph of Figure 18.1

Parameter	*Estimate*	*s. e.*	*t*
Edge	−1.2558	1.3561	0.047
2-Star	−0.0451	0.3551	0.060
3-Star	−0.1084	0.0974	0.073
Triangle	1.4438	0.2073	0.058

Table 18.3 Goodness of Fit for Markov Model for the Mutual Friendship Network (Figure 18.1)

		Simulated		
	Observed	*Mean*	*Std dev*	*t*
Edges	34	34.07	8.91	−0.0077
2-Stars	139	138.35	77.45	0.0084
3-Stars	181	178.05	156.49	0.0189
Triangles	30	29.34	28.39	0.0231
Std dev degrees	2.09	1.71	0.46	0.8296
Skew degrees	0.08	−0.26	0.62	0.5480
Global clustering	0.65	0.54	0.21	0.5246
Mean local clustering	0.64	0.47	0.19	0.4917
Variance local clustering	0.14	0.10	0.04	0.9899

The observed graph, in other words, exhibits levels of clustering and degree heterogeneity that fall within the envelope of values expected for the model. The 1st, 2nd, and 3rd quartiles of the observed geodesic distribution are 1, 3, and 4, respectively; the median values for the distribution of these quartiles across simulations were 2, 2, and 4, suggesting that the model is associated with somewhat more homogeneous internode distances than the data. In Figure 18.8, the distributions of the number of edges, 2-stars, 3-stars, and triangles for the Markov random graph model with these parameter values are shown. While the mean of each distribution is close to the observed value for the Figure 18.1 graph, as expected, it can be seen that the distributions are positively skewed. Indeed, Figure 18.7 shows the impact on one of these statistics—the number of triangles—of changing its corresponding parameter value τ while holding all other parameters constant. It can be seen from Figure 18.7 that the estimated value of 1.4438 is very close to the point of transition between low and high values of the triangle statistic and the positively skewed distribution of the triangle statistic is consistent with the estimated value of τ being just below this point.

Related Model Parameters

In the Markov model just fitted, parameters for 2-stars and 3-stars were included, but parameters for higher-order stars (4-stars, 5-stars, and so on) were assumed to be 0. Arguably, fitting higher-order star parameters might be desirable, because more star parameters will lead to better characterizations of the degree distribution for the network. Indeed, Snijders, Pattison, Robins, and Handcock (2006) proposed that all star parameters be used, but they also imposed a hypothesis about the relationships among stars parameters. Specifically, they assumed that

$$\sigma_{k+1} = (-1/\lambda)\sigma_k, \quad \text{for } k > 1 \text{ and } \lambda \geqslant 1 \text{ a (fixed) constant,}$$

a hypothesis they termed *the alternating k-star hypothesis*. It follows from this hypothesis that

$$\sum_k \sigma_k S_k(\mathbf{x}) = \left[\sum_k (-1)^k S_k(\mathbf{x})/\lambda^{k-2}\right] \sigma_2$$

and, hence, that the entire set of starlike terms in the model can be captured by a single star parameter (σ_2) with a single *alternating k-star* statistic:

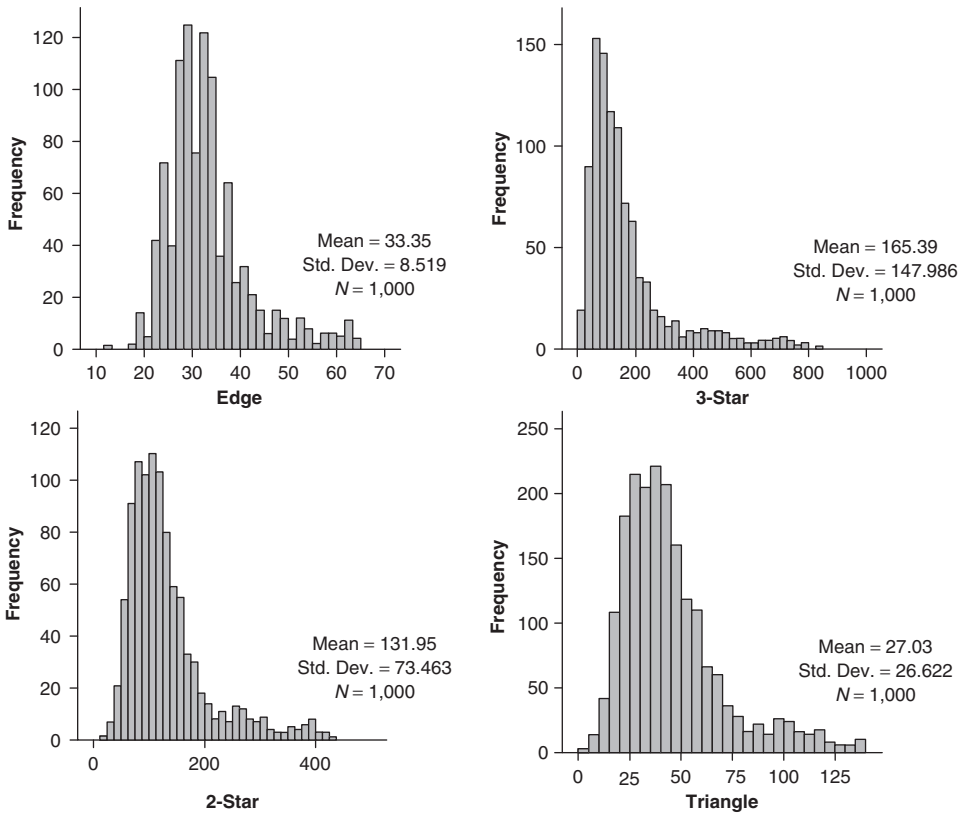

Figure 18.8 Distribution of edge, 2-star, 3-star, and triangle statistics in the Markov random graph distribution with parameters $\theta = -1.2558$, $\sigma_2 = -0.0451$, $\sigma_3 = -0.1084$, $\tau = 1.4438$.

$$S^{[\lambda]}(\mathbf{x}) = \sum_k (-1)^k S_k(\mathbf{x})/\lambda^{k-2}$$

It is of course an empirical matter whether this is an appropriate hypothesis to make. This expression can be simplified to yield a simpler form of the alternating k-star statistic:

$$S^{[\lambda]}(\mathbf{x}) = \lambda^2 \sum_i \{(1 - 1/\lambda)^{k(i)} + k(i)/\lambda - 1\},$$

recalling that $k(i)$ denotes the degree of node i.[3]

Fitting the model

$$\Pr(\mathbf{X} = \mathbf{x}) = \exp(\theta L(\mathbf{x}) + \sigma_2 S^{[\lambda]}(\mathbf{x}))/\kappa$$

to the friendship network in Figure 18.1 yields estimates (standard errors) of -2.7454 (1.3794) and 0.4822 (0.4098) for θ and σ_2, respectively (as before, θ is an edge parameter, and $L(\mathbf{x})$, the number of edges in the network $\mathbf{x}$, is its sufficient statistic). Although this model does a reasonable job in reproducing the standard deviation and the skewness coefficient for the degree distribution, not surprisingly, it does a poor job in recovering network clustering. We present a better-fitting model below.

In some applications to date (e.g., Goodreau, 2007; Snijders et al., 2006), a fixed value of λ (such as 2) has been assumed; Hunter and Handcock (2006) have shown that λ can be treated as a variable within a curved exponential family model, and they have developed an associated estimation method.

[3]Hunter and Handcock (2006) proposed an alternative statistic based on geometrically weighted degree statistics; the resulting model is equivalent, provided that the edge parameter is included.

REALIZATION-DEPENDENT MODELS

A critique of the Markov dependence assumption led Pattison and Robins (2002) to construct a more general class of "realization-dependent" network models. They argued that conditional dependencies among tie variables may emerge from the network processes themselves, with new dependencies created as network ties are generated. For instance, X_{ij} and X_{kl} might become conditionally dependent *if* there is an observed tie between, say, j and k. Baddeley and Möller (1989) termed such models *realization dependent*.

The 4-Cycle Hypothesis

Snijders et al. (2006) argued that in addition to the Markov assumption, two network ties, X_{ij} and X_{kl}, might be conditionally dependent in the case where there is an observed tie between, say, j and k and between l and i; that is, if the presence of a tie from i to j and from k to l would create a 4-cycle in the graph. The rationale for this assumption is that a 4-cycle is a closed structure that can sustain mutual social monitoring and influence, as well as levels of trustworthiness within which obligations and expectations might proliferate (e.g., Coleman, 1988).

Snijders et al. (2006) showed that this assumption led to additional nonzero parameters in an exponential random graph model, including those referring to collections of 2-paths with common starting and ending nodes and collections of triangles with a common base (see Figure 18.9). We define a *k-2-path* to be a subgraph comprising two nodes, i and j, and a set of k paths of length 2 from i to j through distinct intermediate nodes $m_1, m_2, \ldots, m_k$. A *k-triangle* is a subgraph comprising two *connected nodes*, i and j, and a set of k paths of length 2 from i to j through distinct intermediate nodes $m_1, m_2, \ldots, m_k$. If we let ν_k be the model parameter associated with a k-2-path and τ_k the parameter associated with a k-triangle, we can entertain assumptions about the relationships among related parameters (as in the case of k-stars earlier)—namely,

$$\nu_{k+1} = -\nu_k/\lambda$$

and

$$\tau_{k+1} = -\tau_k/\lambda.$$

As for the star parameters, this is just a hypothesis, and its adequacy needs to be assessed. Under this assumption, the statistics

$$U^{[\lambda]}(\mathbf{x}) = \sum_k (-1)^k U_k(\mathbf{x})/\lambda^{k-2}$$

and

$$T^{[\lambda]}(\mathbf{x}) = \sum_k (-1)^k T_k(\mathbf{x})/\lambda^{k-2}$$

become single statistics associated with the parameters ν_1 and τ_1, respectively, where $U_k(\mathbf{x})$ and $T_k(\mathbf{x})$ are the number of k-2-paths and k-triangles in the network $\mathbf{x}$. It should be noted that the value of λ need not be the same for each statistic; as before, Hunter and Handcock have shown how to estimate these parameters.

The parameter estimates presented in Table 18.4 are for a model fitted to the mutual friendship network of Figure 18.1. The positive τ_1 estimate suggests that networks with relatively many triangles are more likely, other statistics being equal, with the cumulative impact of multiple triangles with a common base pair of nodes diminishing as the number of such triangles increases. Likewise, the negative ν_1 estimate suggests that networks with relatively few 2-paths among a pair of nonconnected nodes are more likely, other statistics being equal. Both of these effects are consistent with a pressure toward closure for mutual friendship ties.

The goodness of fit for this model is summarized in Table 18.5. The median values of the quartiles of the geodesic distribution for the random graph distribution simulated from the parameter estimates in Table 18.4 are 2, 3, and 5, suggesting better recovery of short distances than the Markov model, though not of longer ones. Overall, the model of Table 18.4 appears to do a reasonably good job of characterizing the features of the mutual friendship network.

Directed Graph Models

The derivation of similar classes of models for directed graphs is, in principle, very similar to the derivation of models for their nondirected counterparts. Directed graphs give rise, however, to substantially more complicated parameterizations, as a comparison between triadic

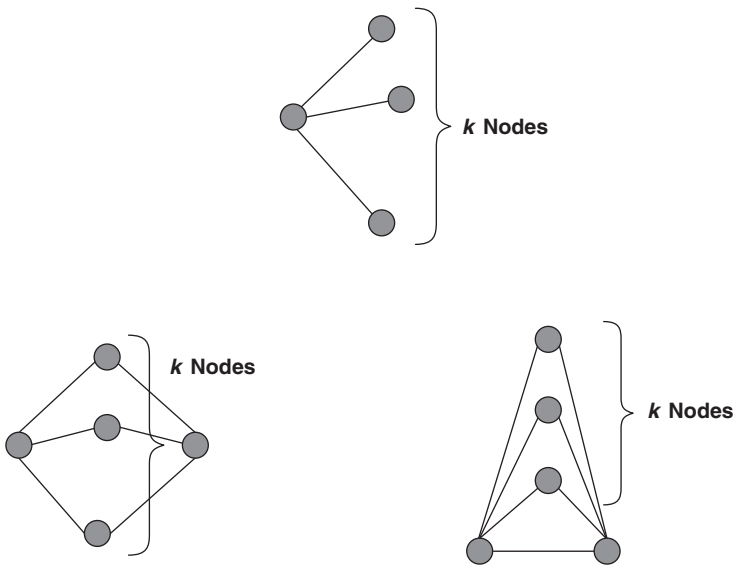

Figure 18.9 The *k*-star, *k*-2-path, and *k*-triangle.

Table 18.4 MCMCMLEs for Realization-Dependent Exponential Random Graph Model for the Mutual Friendship Network (Figure 18.1)

Parameter	*estimate*	*s.e.*	*t*
Edge	−0.0354	1.7851	−0.037
2-Star	−0.0520	0.1094	−0.038
k-Star	0.0674	0.8689	−0.040
k-Triangles	0.7250	0.3159	−0.043
k-2-Paths	−0.5583	0.1727	−0.025

Table 18.5 Goodness of Fit of Realization-Dependent Model for the Mutual Friendship Network

		Simulated		
Statistic	*Observed*	*Mean*	*Std dev*	*t*
Edges	34	34.44	9.33	−0.047
2-Stars	139	145.13	94.54	−0.065
k-Stars	77.8	79.73	34.68	−0.054
k-Triangles	46.0	47.01	28.39	0.023
k-2-Paths	83.3	85.30	30.68	−0.064
Std dev degrees	2.09	1.77	0.60	0.522
Skew degrees	0.08	−0.17	0.46	−0.548
Global clustering	0.65	0.55	0.14	0.696
Mean local clustering	0.64	0.50	0.15	0.439
Variance local clustering	0.14	0.09	0.04	1.207

forms in graphs and directed graphs quickly suggests. There are, as a result, some subtleties to the development of models in the directed graph case (see Robins, Pattison, & Wang, 2006, for further details).

Exogenous Covariates

The general modeling framework can readily accommodate covariates at the node or dyad level, and, of course, if such covariates are regarded as important influences on network tie formation, then they should be included in models for the network. For example, a general and systematic approach to the inclusion of node-level covariates has been outlined by Robins, Elliott, and Pattison (2001), who extended the dependence graph formulation described earlier to include directed dependence relationships from exogenous node-level variables to endogenous tie variables. These developments offer an important means for exploring a wide range of interesting interactions among actor-level and network tie variables, including *homophily* effects (e.g., McPherson, Smith-Lovin, & Cook, 2001) and the effects of spatial locations (e.g., Butts, 2003; Wong, Pattison, & Robins, 2006).

EXTENSIONS

There are a variety of ways in which the models just described may be extended to incorporate richer data forms, including multiple networks, longitudinal data, and changing node and tie sets. In addition, some initial progress has been made on the problem of dealing with missing data. We do not have space for a full account of these interesting and important developments but point to some key developments in each case.

The development of probabilistic models for graphs and directed graphs has been loosely shadowed by the construction of a parallel, albeit generally later, set of models for multiple networks measured on a common node set. Multivariate exponential random graph models are described by Pattison and Wasserman (1999); see also Koehly and Pattison (2005).

Although networks are often measured at a single point in time, they are, in reality, dynamic entities, and there is considerable interest in the processes that underpin their evolution. An important line of work has developed continuous-time Markov process models for network evolution (e.g., see Snijders, 2001). More recently, this framework has been extended to accommodate the possibility of co-evolutionary mechanisms by which network tie change depends on and contributes to change in node attributes (Snijders, Steglich, & Schweinberger, 2007).

CONCLUSION

The field of probabilistic network theory has progressed rapidly in the past 10 years, and as Goodreau (2007) and Robins et al. (2007) have cogently demonstrated, it is now possible to build plausible models for many small and large social networks. Undoubtedly, experience with the current generation of realization-dependent network models will lead to further improvements in model specification and a clearer understanding of how the content of network ties and the contexts in which they are observed might inform model building. Perhaps most important though, the field has now advanced to the point where the promise of a step-change in our understanding of social processes on networks and their consequences might be realized.

ACKNOWLEDGMENTS

We are grateful to Peng Wang and Galina Daraganova for helpful comments on this chapter.

REFERENCES

Albert, R., & Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Baddeley, A., & Möller. (1989). Nearest-neighbor Markov point processes and random sets. *International Statistical Review*, *57*, 89–121.

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B*, *36*, 96–127.

Bollobás, B. (1985). *Random graphs*. London: Academic Press.

Bollobás, B. (1998). *Modern graph theory*. New York: Springer.

Brieger, R. (1981). Comment on "An exponential family of probability distributions for directed graphs". *Journal of the American Statistical Association*, *76*, 51–53.

Burda, Z., Jurkiewicz, J., & Krzywicki, A. (2004). Network transitivity and matrix models. *Physical Review E*, *69*, 026106.

Butts, C. (2003). Predictability of large-scale spatially embedded networks. In R. L. Breiger, K. M. Carley, & P. E. Pattison (Eds.), *Social network models: Workshop summary and papers* (pp. 313–323). Washington, DC: National Academies Press.

Coleman, J. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*(Suppl.), S95–S120.

Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, *6*, 290–297.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Freeman, L. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, British Columbia: Booksurge.

Goodreau, S. (2007). Advances in exponential random graph (p*) models applied to a large social network. *Social Networks*, *29*, 231–248.

Handcock, M. (2004). *Assessing degeneracy in statistical models for social networks* (Working Paper No. 39). Center for Statistics in the Social Sciences, University of Washington.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., & Morris, M. (2004). *Statnet manual for R* (Tech. Rep.). Center for Statistics in the Social Sciences, University of Washington.

Handcock, M., Raftery, A., & Tantrum, J. (2005). *Model-based clustering for social networks* (Working Paper No. 46). Center for Statistics in the Social Sciences, University of Washington.

Hoff, P., Raftery, A., & Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*, 1090–1098.

Holland, P., & Leinhardt, S. (1975). Local structure in social networks. In D. Heise (Ed.), *Sociological methodology 1976* (pp. 1–45). San Francisco: Jossey-Bass.

Holland, P., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*, 33–50.

Hunter, D., & Handcock, M. (2006). Inference in curved exponential families for networks. *Journal of Computational and Graphical Statistics*, *15*, 565–583.

Koehly, L., & Pattison, P. (2005). Random graph models for social networks: Multiple relations or multiple raters. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 162–191). New York: Cambridge University Press.

Kretzschmar, M., & Morris, M. (1996). Measures of concurrency in networks and the spread of infectious disease: The AIDS example. *Social Science and Medicine*, *21*, 1203–1216.

Lorrain, F., & White, H. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, *1*, 49–80.

McDonald, J., Smith, P., & Forster, J. (2007). Markov chain Monte Carlo exact inference for analysing social networks data. *Social Networks*, *29*(1), 127–136.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chkovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, *298*, 824–827.

Nowicki, K., & Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*, 1077–1087.

Park, J., & Newman, M. (2004). Solution of the 2-star model of a network. *Physical Review E*, *70*, 066146.

Pattison, P., & Robins, G. (2002). Neighbourhood-based models for social networks. *Sociological Methodology*, *32*, 301–337.

Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, *52*, 169–193.

Pattison, P., Wasserman, S., Robins, G., & Kanfer, A. (2000). Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology*, *44*, 563–568.

Rapoport, A. (1957). Contributions to the theory of random and biased nets. *Bulletin of Mathematical Biophysics*, *19*, 257–277.

Rapoport, A., & Horvath, W. (1961). A study of a large sociogram. *Behavioral Science*, *6*, 279–291.

Robins, G., Elliott, P., & Pattison, P. (2001). Network models for social selection processes. *Social Networks*, *23*, 1–30.

Robins, G., Pattison, P., & Wang, P. (2006). *Closure, connectivity and degrees: New specifications for exponential random graph models for directed social networks.* University of Melbourne, Melbourne, Australia.

Robins, G., Pattison, P., & Woolcock, J. (2005). Small and other worlds: Global network structures from local processes. *American Journal of Sociology*, *110*, 894–936.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007). Recent developments in exponential random graph (p*) models. *Social Networks*, *29*, 192–215.

Schweinberger, M., & Snijders, T. (2003). Settings in social networks: A measurement model. *Sociological Methodology*, *33*, 307–341.

Snijders, T. (1991). Enumeration and simulation models for 0–1 matrices with given marginals. *Psychometrika*, *56*, 397–417.

Snijders, T. (2001). The statistical evaluation of social network dynamics. In M. Sobel & M. Becker (Eds.), *Sociological methodology 2001* (pp. 361–395). Boston: Basil Blackwell.

Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models [electronic version]. *Journal of Social Structure*, *3*(2). Available from http://www.cmu.edu/joss/content/articles/volume3/Snijders.pdf.

Snijders, T., Pattison, P., Robins, G., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153.

Snijders, T., Steglich, C., & Schweinberger, M. (2007). Modeling the co-evolution of networks and behavior. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 41–71). Mahwah, NJ: Lawrence Erlbaum.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, *28*, 513–527.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, *85*, 204–212.

Wang, P., Robins, G., & Pattison, P. (2006). *Pnet: Program for the estimation and simulation of p* exponential random graph models [User manual]*. Melbourne, Australia: University of Melbourne, Department of Psychology.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks, I. An introduction to Markov graphs and p*. *Psychometrika*, *61*, 401–425.

Wong, L., Pattison, P., & Robins, G. (2006). A spatial model for social networks. *Physica A: Statistical Mechanics and Its Applications*, *360*, 99–120.