THREE

GRAPH-THEORETIC AND STATISTICAL MODELS

3.1 INTRODUCTION

In this chapter, we consider several graph-theoretic and probabilistic models for a social network, which we do under different assumptions related to two basic parameters stated earlier in Chapter 1—namely, the number of vertices (n) and the number of arcs (m). We will take a social network to be a digraph. Models analogous to some of these can also be considered for graphs and weighted digraphs, although we will not discuss them in detail.

At the outset, we mention that when we talk of models in this chapter, we do not imply that any of them is a typical realistic representation of the situation in real life. We are not trying to build or present such models. Rather, the models we present can be used as a sort of *null model* with which one can standardize some of the parameters or statistics in the underlying social networks.

Each model generally stipulates in some way the set of all possible digraphs of which the observed digraph is an element. A *statistical model*, moreover, assigns a probability distribution, depending on some parameters, over the class of all possible digraphs. For some purposes, one may use a probability model that does not completely specify the probability of each possible digraph. By a *graph-theoretic model*, we simply mean a model that uses some digraph parameters and is not probabilistic.

The general procedure we adopt for obtaining a standardized measure of any characteristic of a social network is as follows. We start with an initial simple real valued measure X, which is a function of the observed network, for the characteristic under consideration. We then choose an appropriate model and standardize X as follows.

If the model is statistical, we take $P(X \leq x)$ to be the standardized measure, where x is the value of X for the particular network observed. This measure lies between 0 and 1 and can be converted to a percentage by multiplying by 100. In case the distribution of X is not known, one can still use $(x - E(X))/\sigma(X)$ as a measure and can get some idea about the tail probability from Chebychev's inequality.

If the model is graphical, we find the minimum $x_{\min}$ and the maximum $x_{\max}$ of the values X takes. Then we scale the observed value x in its range and take

$$\frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

as the standardized measure. The situation is sometimes complicated by the fact that not all (integer) values between $x_{\min}$ and $x_{\max}$ are attained by X. Then, perhaps one can choose one of the following alternatives: (1) use the above measure regardless of whether X takes all values between $x_{\min}$ and $x_{\max}$ and (2) use the probabilistic measure assuming that the distribution of X gives equal probabilities to each of the values taken by X.

Thus, under each of the graphical models, we want to find the range (or at least the minimum and the maximum values) of the crude measure X. Under each of the probabilistic models, we want to find the distribution of X, either exactly or at least approximately, and, if this is not possible, then the mean and the variance of X. We consider the following variables X, some of which will be used in the next chapter to construct measures of various characteristics of a social network:

- the number of arcs;
- the out-degree and the in-degree of a vertex;
- the maximum out-degree and the maximum in-degree in the digraph;
- the numbers of sources, sinks, and isolates;

- the number of symmetric pairs;
- the diameter;
- the radius;
- the number of reachable pairs;
- the number of pairs reachable in k steps ($k = 2$ or 3);
- the numbers p and q of strong and weak components, $p - q$;
- the number of arcs within strong components;
- the number h of arcs between strong components;
- the number of arcs uv such that $d(v, u) \leq 2$; and
- the clique number.

Under the probabilistic models, we will also look at the following, where G denotes a random digraph.

- P(G is symmetric),
- P(G is asymmetric),
- P(G is complete asymmetric),
- P(G is a tree),
- P(G is acyclic),
- P(G is strongly connected), and
- P(G is weakly connected).

To avoid trivialities, we will assume that $n \geq 2$ in all the models considered below, although we allow m to be 0. Also note that the probability distribution of the sources, sinks, and isolates has received special attention as these immediately reveal some important features of social structures. This is discussed later in Chapter 5.

It is perhaps worth mentioning that, in applying these to actual social networks, one has to modify some of the above variables X for various reasons. First, X may not precisely measure the characteristic it is intended to measure (often the latter has different versions or nuances, all of which cannot be captured by a single variable). Then, the value of X itself may be very difficult to find even for the observed network, not to mention the distribution of X, especially if the number of vertices is large, although sometimes, using various computational and theoretical techniques, it may be possible to compute the value for reasonably sized networks. For many of the variables, the exact range or even the values of $x_{\min}$ and $x_{\max}$ and the distribution or even the mean and variance of X are not known exactly, and only occasionally some theoretical approximations or bounds are known. Sometimes it may be possible to use iterative methods or simulation to get approximations, and one has to be satisfied with them. Finally, many of the definitions or properties of graph-theoretic variables are too stringent to be of use in real-life situations, and one has to either modify the variable or use some sort of cutoff to decide whether one can consider a network to have the particular property.

It may not be out of context to also mention that discussions on specific graph parameters such as sinks, sources, and isolates may be overemphasized in the later parts of this chapter. These terms have natural sociological interpretations, as discussed in Chapter 5. The theoretical study of such parameters requires probabilistic arguments, and we thought it proper to derive some results and present them hereinafter along with the description of other more useful parameters such as out-degree, in-degree, and reciprocity.

The basic assumptions underlying different models that will be considered in this chapter are summarized in the following table. We have conceptualized four categories of models (I–IV), penetrating step-by-step according to four different levels of available information related to the formation of the social network. Minimally, since the level of information available is that of only the size of a network (i.e., the number of actors (n)), we begin with n as given. Therefore, the possible digraphs are all digraphs with vertex set $v_1, v_2, \ldots\ldots, v_n$ with no additional assumptions. Models I.1 to I.3 fall under this category. At the next level, the quantum of interaction among the n actors (i.e., total number of ties of interaction or arcs (m)) is also assumed to be known. Models II.1 to II.2 fall in this category. Information available on the out-degrees of the actors—that is, $(d_1, d_2, \ldots\ldots, d_n)$—gives us the third level of modeling, and III.1 to III.3 deal with this aspect. Models IV.1 to IV.2 deal with the situation when both the out-degree and in-degree sequences are known. For each of the four categories, we have considered graph-theoretic (deterministic) and statistical (probabilistic) versions along with appropriate special cases, if any. The remaining models (V–VII) are known probabilistic models that can be related to those stated in Categories II and III above.

Model assumptions

Model no.	*Assumptions*
I.1	None
I.2	All digraphs are equally likely
I.3	$P(v_i v_j \text{ is an arc}) = p$ for $i \neq j$, distinct pairs being independent
II.1	The number of edges is m
II.2	The number of edges is m and all possible digraphs are equally likely
III.1	$d^+(v_i) = d_i$ for $i = 1, 2, \ldots, n$
III.2	$d^+(v_i) = d_i$ for $i = 1, 2, \ldots, n$ and all possible digraphs are equally likely
III.3	$\|N(v_i)\| = d_i$, $N(v_i) \subseteq P_i$ for $i = 1, 2, \ldots, n$ and all possible digraphs are equally likely
IV.1	$d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$
IV.2	$d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$ and all possible digraphs are equally likely
V	$P(v_j v_i \in A \mid v_i v_j \in A) = P(v_j v_i \in A) + \tau P(v_j v_i \notin A)$ and $P(v_i v_j \in A) = d_i/(n-1)$ whenever $i \neq j$
VI	$P((0, 0)) = \exp(\lambda_{ij})$; $P((1, 0)) = \exp(\lambda_{ij} + \theta + \alpha_i + \beta_j)$; $P((0, 1)) = \exp(\lambda_{ij} + \theta + \alpha_j + \beta_i)$; $P((1, 1)) = \exp(\lambda_{ij} + 2\theta + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho)$; for an ordered pair of vertices (i, j) in a digraph G with the scores for dyadic movements expressed by $2^2 = 4$ combinations of 0s and 1s; $\sum \alpha_i = 0$ and $\sum \beta_i = 0$
VII	$P(G) = \text{const.} \exp(\theta m + \rho s + \sum \alpha_i d_i + \sum \beta_i e_i)$ for any digraph G; $\sum \alpha_i = 0$ and $\sum \beta_i = 0$

3.2 MODELS FIXING THE TOTAL NUMBER OF VERTICES

Model I.1

This is the simplest graphical model one can consider and takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ comprising n vertices as fixed and assumes that all the $2^{n(n-1)}$ digraphs on V are *actually* possible.

Clearly, the range of the number m of arcs is $\{0, 1, \ldots, n(n-1)\}$.

The range of the out-degree d_i (as well as the in-degree e_i) of the ith vertex is $\{0, 1, \ldots, n-1\}$ for each i. The range of $d_{\max}$, defined as $\max(d_1, d_2, \ldots, d_n)$, as well as that of $e_{\max}$, defined as $\max(e_1, e_2, \ldots, e_n)$, is also $\{0, 1, \ldots, n-1\}$.

The range of the number of sources (as well as the number of sinks) is $\{0, 1, \ldots, n\}$. However, the range of the number of isolates is $\{0, 1, \ldots, n-2, n\}$ (note that if $n-1$ of the vertices are isolates, the remaining vertex is automatically an isolate).

The range of the number s of symmetric pairs is $\{0, 1, \ldots, n(n-1)/2\}$.

The range of the diameter is $\{1, 2, \ldots, n-1, \infty\}$. (To get a digraph with diameter $k < \infty$, take a complete symmetric digraph on $n-k+1$ vertices and attach a symmetric path on k vertices at some vertex.) The range of the radius is $\{1, 2, \ldots, n-1, \infty\}$. (To get a digraph with radius $k < \infty$, take a directed path on $k+1$ vertices and join the first vertex *to* the remaining $n-k-1$ vertices; then, the first vertex of the path is a center, and the radius is k.)

Next we study R, the number of reachable pairs of *distinct* vertices. (Note that we are leaving out pairs of the type (u, u), even though u is reachable from u in a trivial sense.) Clearly, the minimum and maximum values of R are 0 and $n(n-1)$, but the range S_n of R is not continuous. It can be shown by induction on n that $S_n \subseteq T_n$, where

$$T_n = \{0, 1, \ldots, n^2 - 3n + 4\} \cup \{n^2 - 2n + 1, n^2 - n\}. \tag{3.2}$$

We give S_n below for n up to 15 for ready reference.

$$
\begin{aligned}
S_1 &= \{0\} \subseteq T_1,\\
S_2 &= \{0, 1, 2\} = T_2,\\
S_3 &= \{0, 1, 2, 3, 4, 6\} = T_3,\\
S_4 &= \{0, 1, \ldots, 9, 12\} = T_4,\\
S_5 &= \{0, 1, \ldots, 14, 16, 20\} = T_5,\\
S_6 &= \{0, 1, \ldots, 22, 25, 30\} = T_6,\\
S_7 &= \{0, 1, \ldots, 28, 30, 31, 32, 36, 42\} = T_7 - \{29\},\\
S_8 &= \{0, 1, \ldots, 44, 49, 56\} = T_8,\\
S_9 &= \{0, 1, \ldots, 52, 54, 56, 57, 58, 64, 72\} = T_9 - \{53, 55\},\\
S_{10} &= \{0, 1, \ldots, 67, 69, 72, 73, 74, 81, 90\} = T_{10} - \{68, 70, 71\},\\
S_{11} &= \{0, 1, \ldots, 84, 86, 90, 91, 92, 100, 110\} = T_{11} - \{85, 87, 88, 89\},\\
S_{12} &= \{0, 1, \ldots, 97, 99, \ldots, 103, 105, 110, 111, 112, 121, 132\},\\
S_{13} &= \{0, 1, \ldots, 117, 120, \ldots, 124, 126, 132, 133, 134, 144, 156\},\\
S_{14} &= \{0, 1, \ldots, 139, 142, \ldots, 147, 149, 156, 157, 158, 169, 182\},\\
S_{15} &= \{0, 1, \ldots, 163, 166, 168, \ldots, 172, 174, 182, 183, 184, 196, 210\}.
\end{aligned}
$$

Moreover, Rao (2002) has shown that for $n \le 208$,

$$x \in S_n \text{ if and only if } x - k(n-1) \in S_{n-k} \text{ for some } k \text{ with } 1 \le k \le n-1, \quad (3.3)$$

using which S_n can be determined for $n \le 208$. He also showed that if $f(n)$ is defined by $\{0, 1, \ldots, f(n)\} \subseteq S_n$ and $f(n)+1 \notin S_n$, then $f(n) \ge (n - \lfloor n^{0.57} \rfloor)(n-1)$ and that this bound is fairly good. It has also been found empirically that the number of elements in S_n is close to $(n - \lfloor n^{0.45} \rfloor)(n-1)$ at least for $n \le 208$.

Until now, we considered the number of pairs reachable in an arbitrary number of steps. Let $R^{(k)}(G)$ denote the number of pairs (u, v) of distinct vertices in G such that v is reachable in k or fewer steps from u, and let $S_n^{(k)}$ be the range of $R^{(k)}(G)$ as G varies over all networks on n vertices. Rao (2002) has shown that $S_n^{(2)} = S_n$ for $n = 1, 2, 3$ and $S_n^{(2)} = \{0, 1, \ldots, n(n-1)\}$ whenever $n \ge 4$; $S_n^{(3)} = S_n$ for $n = 1, 2, 3, 4$ and $S_n^{(3)} = \{0, 1, \ldots, n(n-1)\}$ whenever $n \ge 5$. He has also shown that for every $k \ge 2$, $S_n^{(k)} = \{0, 1, \ldots, n(n-1)\}$ provided, $n \ge k + \lfloor (k+1)^{0.57} \rfloor + 2$.

The range is $\{1, 2, \ldots, n\}$ for the number of strong components, the number of weak components, and the clique number.

It is easy to see that the range of the difference $p - q$ between the number of strong components p and the number of weak components q is also $\{0, 1, \ldots, n-1\}$. The range of the number h of arcs between strong components is $\{0, 1, \ldots, \binom{n}{2}\}$. This is because such arcs cannot be reciprocated. The range of the minimum number P of paths, formed with arcs joining different strong components of G and covering the vertex set, is clearly $\{1, 2, \ldots, n\}$.

Model I.2

Model I.2 is a probabilistic version of Model I.1, and takes the vertex set as fixed (e.g., $V = \{v_1, v_2, \ldots, v_n\}$) and assumes that all the $2^{n(n-1)}$ possible digraphs on V are equally likely.

Let X_{ij} be the random variable taking value 1 if $v_i v_j$ is an arc and 0 otherwise. Then under the present model, $P(X_{ij} = 1) = 1/2$ since there are $2^{n(n-1)-1}$ digraphs with $v_i v_j$ as an arc. Also, it is easy to see that the $n(n-1)$ random variables X_{ij} $(1 \le i \ne j \le n)$ are mutually independent. Thus, the model is equivalent to $v_i v_j$, which is an arc with probability $1/2$, and the ordered pairs are all mutually independent. Hence, a random digraph G under the model can be generated by making $v_i v_j$ an arc with probability $1/2$ for each ordered pair (i, j), with distinct ordered pairs being independent. Repeating this, one can generate any given number of random digraphs and estimate the distribution of any statistic under the model by simulation.

It is easy to see that under the present model, the number of arcs m has the binomial distribution $B(n(n-1), 1/2)$ since m is the sum of the $n(n-1)$ independent Bernoulli random variables X_{ij}. The notation $B(..)$ will be used for a binomial distribution without any further explanation.

Since the out-degree d_i of v_i is $\sum_{j \neq i} X_{ij}$, it follows that d_i has the binomial distribution $B(n-1, 1/2)$. Hence, $E(d_i) = (n-1)/2$ and $V(d_i) = (n-1)/4$. Since the d_is are independent, the distribution of $d_{\max} := \max(d_1, d_2, \ldots, d_n)$ can be computed easily, although one cannot give a closed formula for it. Note that $P(d_{\max} \leq k) = (P(d_1 \leq k))^n$. It is easy to see that the in-degree e_i of v_i also has the distribution $B(n-1, 1/2)$, and $d_{\max}$ and $e_{\max}$ have the same distribution.

To give an example, let $n = 3$. Then each d_i takes values 0, 1, and 2 with probabilities $1/4$, $1/2$, and $1/4$. Hence, we have $P(d_{\max} = 0) = (1/4)^3 = 1/64$. Also, $P(d_{\max} \leq 1) = (3/4)^3 = 27/64$, so $P(d_{\max} = 1) = 13/32$ and $P(d_{\max} = 2) = 37/64$. When $n = 4$, it can be checked that $d_{\max}$ takes values 0, 1, 2, and 3 with probabilities 1/4,096, 255/4,096, 2,145/4,096, and 1,695/4,096, respectively.

The probability that v_i is a source is $1/2^{n-1}$. Also, different v_is being sources are independent events, so the number of sources has the distribution $B(n, 1/2^{n-1})$. It follows similarly that the probability that v_i is a sink is also $1/2^{n-1}$, and the number of sinks has the distribution $B(n, 1/2^{n-1})$.

The probability that v_i is an isolated vertex is $1/2^{2n-2}$. But the events in which different vertices are isolates are not independent. For example, if any $n-1$ vertices are isolates, it follows that the remaining vertex is also an isolate. So, to find the distribution of the number of isolates, we use the following formulae (see Feller, 1968). The probability that exactly k of the events $A_1, A_2, \ldots, A_n$ occurs is

$$S_k - \binom{k+1}{k} S_{k+1} + \binom{k+2}{k} S_{k+2} - + \cdots + (-1)^{n-k} \binom{n}{k} S_n, \qquad (3.4)$$

and the probability that at least k of $A_1, A_2, \ldots, A_n$ occurs is

$$S_k - \binom{k}{k-1} S_{k+1} + \binom{k+1}{k-1} S_{k+2} - + \cdots + (-1)^{n-k} \binom{n-1}{k-1} S_n, \qquad (3.5)$$

where S_k denotes $\sum P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k})$, with the sum being taken over all $i_1, i_2, \ldots, i_k$ such that $1 \le i_1 < i_2 < \cdots < i_k \le n$. Clearly, now the probability that k given vertices are isolates is

$$\frac{1}{2^{k(k-1)+2k(n-k)}} = \frac{1}{2^{k(2n-k-1)}}.$$

Hence, taking A_i to be the event that the ith vertex is an isolate, we see that the probability that there are exactly k isolates is

$$\binom{n}{k}\frac{1}{2^{k(2n-k-1)}} - \binom{k+1}{k}\binom{n}{k+1}\frac{1}{2^{(k+1)(2n-k-2)}} + \binom{k+2}{k}$$
$$+\binom{n}{k+2}\frac{1}{2^{(k+2)(2n-k-3)}} - + \cdots + (-1)^{n-k}\binom{n}{k}\frac{1}{2^{n(n-1)}}. \quad (3.6)$$

The probability that there are at least k isolates can be found by using (3.2). Taking the events that different vertices are isolates to be nearly independent, we see that the distribution of the number of isolates is approximately $B(n, 1/2^{2n-2})$. However, this is not of much importance as the probability that there is no isolate is more than 0.999 for all $n \ge 10$.

By definition, the probability that G is any particular digraph (including the null digraph and the complete symmetric digraph) is $1/2^{n(n-1)}$.

The probability that G is symmetric is $1/2^{n(n-1)/2}$ since, for G to be symmetric, either none or both of $v_i v_j$ and $v_j v_i$ should be arcs for each unordered pair $\{i, j\}$ with $i \ne j$.

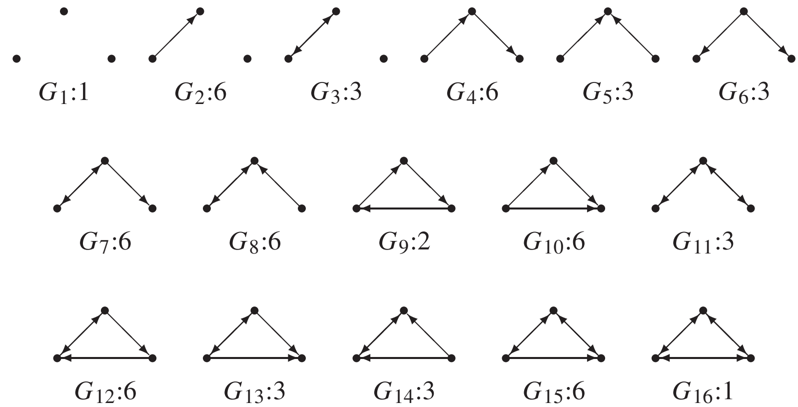
The probability that G is asymmetric and the probability that G is complete are both $(3/4)^{n(n-1)/2}$ since, for G to be asymmetric, at most one of $v_i v_j$ and $v_j v_i$ should be an arc, and for G to be complete, at least one of $v_i v_j$ and $v_j v_i$ should be an arc, for each unordered pair $\{i, j\}$, with $i \ne j$.

The number of symmetric pairs $s(G)$ has the distribution $B(n(n-1)/2, 1/4)$ since $s(G)$ is the sum of the $n(n-1)/2$ independent Bernoulli variables Y_{ij}, where $Y_{ij} = X_{ij} X_{ji}$ for any unordered pair $\{i, j\}$ with $i \ne j$.

It is much more difficult to deal with probabilities of events depending on the distance between vertices because $d(v_i, v_j)$ depends not only on what happens at v_i and v_j but also on what happens in other parts of the digraph.

For example, to find the probability that G is strongly connected, we have to find the number $f(n)$ of strongly connected digraphs on the vertex set $\{v_1, v_2, \ldots, v_n\}$. A method for computing this number $f(n)$ for any n is known (see Harary, 1988). (The problem is usually referred to as enumeration

Figure 3.1



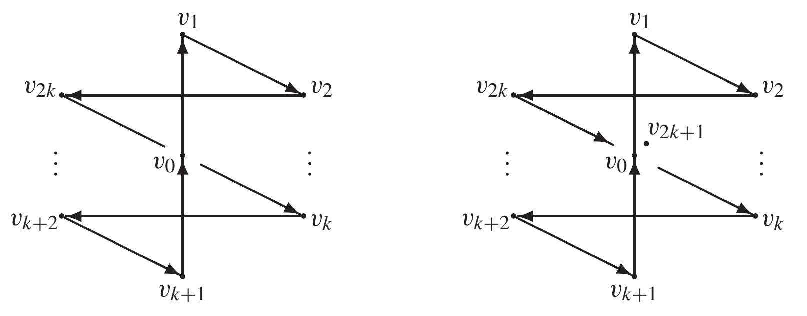
of strongly connected labeled digraphs on n vertices, with the word *labeled* signifying that the vertex set is fixed and we are not counting nonisomorphic digraphs.) However, this involves generating functions, and no closed formula is known for $f(n)$. Thus, even finding the probability that G is strongly connected is difficult. When $n = 2$, this probability is $1/4$. If $n = 3$, out of the 64 possible digraphs, 18 are strongly connected, so the probability is $9/32$. Here, the strongly connected digraphs consist of one digraph with six arcs; six digraphs with five arcs; three digraphs with four arcs forming two reciprocal pairs; six digraphs with four arcs, two of which form a reciprocal pair; and two digraphs with three arcs forming a circuit (see Figure 3.1). When $n = 4$, it can be checked that out of 4,096 possible digraphs, 1,606 are strongly connected, so the probability that G is strongly connected is 1,606/4,096 = 0.392.

When $n = 2$, the probability that G is weakly connected is $3/4$. When $n = 3$, 54 out of the 64 possible digraphs are weakly connected, so the probability is $54/64 = 0.844$. When $n = 4$, it can be checked that out of 4,096 possible digraphs, 3,834 are weakly connected, so the probability is 3,834/4,096 = 0.936.

In Figure 3.2, we give the 16 nonisomorphic digraphs on three vertices. Along with each of these, we also give the number of digraphs on $\{v_1, v_2, v_3\}$ isomorphic to it.

The probability that G has diameter 1 is clearly $1/2^{n(n-1)}$. When $n = 3$, it is easy to check from Figure 3.2 that the probability that G has diameter 2 is $17/64$, and the probability that G has diameter ∞ is $23/32$.

Figure 3.2



Let $n = 3$. Then $P(r(G) = 1) = P(d_{\max} = 2) = 37/64$, where $r(G)$ is the radius of G. It can be checked that 14 digraphs have radius 2 and 13 digraphs have radius ∞, so $P(r(G) = 2) = 7/32$ and $P(r(G)) = \infty) = 13/64$.

We mention that for general n and k, even expressions such as (3.3) are difficult to find for $P(r(G) = k)$ and $P(d(G) = k)$. However, it is easy to prove that $P(d(G) = 2) \to 1$ as $n \to \infty$. To see this, let A_i be the event that at least one of $v_1 v_i$ and $v_i v_2$ is not an arc. Then $P(A_i) = 3/4$ for all $i \neq 1, 2$. Since $P(v_1 v_2 \text{ is not an arc}) = 1/2$, it follows that $P(d(v_1, v_2) > 2) = (3/4)^{n-2}/2$. Since $P(d(v_i, v_j) > 2) = P(d(v_1, v_2) > 2)$, whenever $i \neq j$, we have $P(d(G) > 2) = P(d(v_i, v_j) > 2 \text{ for at least one pair } (i, j)) \leq n(n-1)(3/4)^{n-2}/2 \to 0$ as $n \to \infty$. Since $P(d(G) = 1) = 1/2^{n(n-1)} \to 0$, we have $P(d(G) = 2) \to 1$ as $n \to \infty$.

Since $P(d(G) = 2) \to 1$, we also have $P(r(G) \leq 2) \to 1$ as $n \to \infty$. Now we show that $P(r(G) = 1) \to 0$ as $n \to \infty$. For any fixed i, let E_i be the event that $v_i v_j$ is an arc for all $j \neq i$. Then $P(E_i) = 1/2^{n-1}$. Also, E_is are independent, so

$$
\begin{aligned}
P(r(G) = 1) &= P(E_1 \cup E_2 \cup \cdots \cup E_n) = 1 - P(\bar{E}_1 \cap \bar{E}_2 \cap \cdots \cap \bar{E}_n) \\
&= 1 - \left(1 - \frac{1}{2^{n-1}}\right)^n = \sum_{k=1}^{n} (-1)^{k-1} \frac{\binom{n}{k}}{2^{k(n-1)}} \leq \frac{n}{2^{n-1}} \to 0 \qquad (3.7)
\end{aligned}
$$

as $n \to \infty$. Here the inequality follows from the fact that $|a_k|/|a_{k+1}| > 1$ for all k, where a_k denotes the kth term in the sum (it also follows from the fact that $P(\cup E_i) \leq \sum P(E_i)$). Hence, $P(r(G) = 1) \to 0$, and $P(r(G) = 2) \to 1$ as $n \to \infty$.

The distribution of the clique number $\omega(G)$ is difficult to find for general n, although it can be worked out for small n using (3.2). It may be noted that, for any n, $P(\omega(G)=1)=P(G \text{ is asymmetric})=(3/4)^{n(n-1)/2}$ and $P(\omega(G)=n)=1/2^{n(n-1)}$. Hence, when $n=3$, the clique number takes values 1, 2, and 3 with probabilities 27/64, 9/16, and 1/64. Next let $n=4$. Then $P(\omega(G)=1)=729/4{,}096$, so $P(\omega(G)\geq 2)=3{,}367/4{,}096$. To find $P(\omega(G)\geq 3)$, let A_1, A_2, A_3, A_4 be the events that $\{v_1, v_2, v_3\}$, $\{v_1, v_2, v_4\}$, $\{v_1, v_3, v_4\}$, and $\{v_2, v_3, v_4\}$ induce complete symmetric digraphs. Then $P(A_i)=1/2^6$, $P(A_i \cap A_j)=1/2^{10}$ whenever $i \neq j$, $P(A_i \cap A_j \cap A_k)=1/2^{12}$ whenever i, j, k are distinct, and $P(A_1 \cap A_2 \cap A_3 \cap A_4)=1/2^{12}$. So $P(\omega(G)\geq 3)=4/64-6/1{,}024+4/4{,}096-1/4{,}096=235/4{,}096$. Hence, $P(\omega(G)=2)=3{,}132/4{,}096$, $P(\omega(G)=3)=234/4{,}096$, and $P(\omega(G)=4)=1/4{,}096$. However, if we try to find $P(\omega(G)\geq 3)$ in the same way when $n=5$, we note that $P(A_i \cap A_j \cap A_k)$ depends on what i, j and k are. Thus, the formulae become more complicated as n increases.

Results of Simulation

It should be evident that for some of the statistics, the exact distributional properties are hard to derive. These are analytically intractable unless n is small. For ready reference, therefore, we give the distribution of various statistics considered above, for some values of n. These were obtained by simulation using 100,000 random digraphs (except for small values of n when the exact distribution can be computed). The error in the estimate of any probability should not exceed 0.005 and is expected to be much less (less than 0.001 when the probability is less than 0.01). We do not give the distributions of $m(G)$ and $s(G)$ as these are binomial distributions. Note that a dash in an entry in the table means that the probability is either 0 or is positive but less than 0.0005.

Maximum Out-Degree

n	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
2	.250	.750	–	–	–	–	–	–	–	–
3	.016	.406	.578	–	–	–	–	–	–	–
4	–	.062	.525	.413	–	–	–	–	–	–
5	–	–	.151	.572	.274	–	–	–	–	–
6	–	–	.015	.273	.537	.174	–	–	–	–
10	–	–	–	–	.001	.053	.337	.430	.159	.019

	Maximum Out-Degree (Continuation)							
	11	*12*	*13*	*14*	*15*	*16*	*17*	*18*
$n = 20$	.019	.155	.350	.301	.131	.036	.006	.001

	Maximum Out-Degree (Continuation)										
	22	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*
$n = 40$	.001	.015	.095	.230	.281	.207	.106	.044	.015	.005	.001

	Sources					*Isolates*				
	0	*1*	2	*3*	4	*0*	*1*	2	*3*	4
n										
2	.250	.500	.250	–	–	.750	–	.250	–	–
3	.423	.421	.140	.016	–	.844	.140	–	.016	–
4	.586	.336	.071	.007	–	.943	.052	.005	–	–
5	.725	.240	.033	.002	–	.981	.019	.001	–	–
6	.826	.160	.013	.001	–	.994	.006	–	–	–
10	.981	.019	–	–	–	1	–	–	–	–
20	1	–	–	–	–	1	–	–	–	–
40	1	–	–	–	–	1	–	–	–	–

	Diameter						*Radius*					
	1	*2*	*3*	*4*	*5*	∞	*1*	*2*	*3*	*4*	*5*	∞
n												
2	.250	–	–	–	–	.750	.750	–	–	–	–	.250
3	.016	.267	–	–	–	.717	.578	.219	–	–	–	.203
4	–	.174	.219	–	–	.607	.413	.422	.047	–	–	.118
5	–	.131	.318	.093	–	.458	.274	.609	.054	.007	–	.055
6	–	.116	.413	.136	.020	.316	.174	.752	.045	.007	.001	.021
10	–	.124	.746	.086	.005	.039	.020	.974	.006	–	–	–
20	–	.475	.525	–	–	–	–	1	–	–	–	–
40	–	.987	.013	–	–	–	–	1	–	–	–	–

	Strong Components						*Weak Components*			
	1	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*
n										
2	.250	.750	–	–	–	–	.750	.250	–	–
3	.283	.326	.391	–	–	–	.844	.140	.016	–
4	.393	.285	.188	.134	–	–	.936	.059	.005	–
5	.542	.257	.116	.057	.029	–	.979	.020	.001	–
6	.684	.213	.066	.023	.010	.004	.994	.006	–	–
10	.961	.037	.001	–	–	–	1	–	–	–
20	1	–	–	–	–	–	1	–	–	–

	Clique Number					
	1	*2*	*3*	*4*	*5*	*6*
n						
2	.750	.250	–	–	–	–
3	.422	.562	.016	–	–	–
4	.178	.765	.057	–	–	–
5	.057	.813	.129	.001	–	–
6	.014	.754	.229	.003	–	–
10	–	.263	.694	.043	–	–
20	–	–	.470	.517	.013	–
40	–	–	–	.648	.349	.003

Model I.3

Under Model I.2, the expected number of arcs is $n(n-1)/2$. So when m is very small or very large (close to $n(n-1)$), Model I.2 is not appropriate, and we consider Model I.3.

In this probabilistic model, we again take the vertex set to be fixed and assume that all the $2^{n(n-1)}$ digraphs on V are possible. But we stipulate that $v_i v_j$ is chosen as an arc with a fixed probability p $(0 < p < 1)$ and that the ordered pairs are all mutually independent. Note that Model I.2 is the special case of Model I.3 corresponding to $p = 1/2$.

Let X_{ij} be defined as before. Then, under the present model, we have $P(X_{ij} = 1) = p$, and X_{ij}s are independent. So the number of arcs m has the binomial distribution $B(n(n-1), p)$, and the out-degree d_i as well as the in-degree e_i of v_i has the same binomial distribution $B(n-1, p)$.

Clearly, the probability of getting any particular digraph G_0 on V is $p^m q^{n(n-1)-m}$, where m is the number of arcs in G_0 and $q = 1 - p$. Hence, any two digraphs with the same number of arcs are equally likely; in particular, $P(G = G_0) = P(G = G_0^c)$, where G_0^c denotes the converse of G_0.

Here, the probability that v_i is a source is q^{n-1}. Also, the number of sources has the distribution $B(n, q^{n-1})$. Similarly, the probability that v_i is a sink is q^{n-1}, and the number of sinks has the distribution $B(n, q^{n-1})$.

The probability that v_i is an isolated vertex is q^{2n-2}. The probability that k given vertices are isolates is $q^{k(2n-k-1)}$. Hence, the probability that there are exactly k isolates is

$$\binom{n}{k}q^{k(2n-k-1)} - \binom{k+1}{k}\binom{n}{k+1}q^{(k+1)(2n-k-2)} + \binom{k+2}{k}\binom{n}{k+2}q^{(k+2)(2n-k-3)} - + \cdots + (-1)^{n-k}\binom{n}{k}q^{n(n-1)}. \quad (3.8)$$

Again, an expression, similar to (3.2), for the probability that there are at least k isolates can be written down.

The probability that G is a null digraph is $q^{n(n-1)}$, and the probability that G is the complete symmetric digraph is $p^{n(n-1)}$.

The probability that G is symmetric is $(p^2 + q^2)^{n(n-1)/2}$, the probability that G is asymmetric is $(1 - p^2)^{n(n-1)/2}$, and the probability that G is complete is $(1 - q^2)^{n(n-1)/2}$.

The number of symmetric pairs $s(G)$ has the distribution $B(\binom{n}{2}, p^2)$.

When $n = 2$, the probability that G is strongly connected is p^2. If $n = 3$, the probability is $p^6 + 6p^5q + 9p^4q^2 + 2p^3q^3$. This can be seen easily from Figure 3.1.

Here also expressions for $P(r(G) = k)$ and $P(d(G) = k)$ cannot be found but, rather surprisingly, it is easy to prove that $P(d(G) = 2) \to 1$ and $P(r(G) = 2) \to 1$ as $n \to \infty$, provided only that $0 < p < 1$. To see this, let A_i be the event that at least one of v_1v_i and v_iv_2 is not an arc. Then $P(A_i) = 1 - p^2$ for all $i \neq 1,2$. Since $P(v_1v_2 \text{ is not an arc}) = 1 - p$, we have $P(d(v_1, v_2) > 2) = (1 - p)(1 - p^2)^{n-2}$. It follows as before that $P(d(G) > 2) \leq n(n-1)(1-p)(1 - p^2)^{n-2} \to 0$ as $n \to \infty$. Since $P(d(G) = 1) = p^{n(n-1)} \to 0$, it follows that $P(d(G) = 2) \to 1$ as $n \to \infty$. Hence, we also have $P(r(G) \leq 2) \to 1$ as $n \to \infty$. Now we show that $P(r(G) = 1) \to 0$ as $n \to \infty$. For any fixed i, let E_i be the event that v_iv_j is an arc for all $j \neq i$. Then $P(E_i) = p^{n-1}$. Also, E_is are independent, so

$$\begin{aligned} P(r(G) = 1) &= P(E_1 \cup E_2 \cup \cdots \cup E_n) \\ &= 1 - P(\bar{E}_1 \cap \bar{E}_2 \cap \cdots \cap \bar{E}_n) = 1 - \left(1 - p^{n-1}\right)^n. \quad (3.9) \end{aligned}$$

Moreover, $P(E_1 \cup E_2 \cup \cdots \cup E_n) \leq P(E_1) + P(E_2) + \cdots + P(E_n) = np^{n-1} \to 0$. Hence, $P(r(G) = 1) \to 0$ and $P(r(G) = 2) \to 1$ as $n \to \infty$.

3.3 MODELS FIXING THE TOTAL NUMBER OF VERTICES AND THE TOTAL OF ALL ARCS

Model II.1

In Model I.1, we took only the number of vertices (n) as given and assumed that all values of m from 0 to $n(n-1)$ are actually possible. Often this is not realistic, particularly when n is large, so we introduce Model II.1.

This graphical (nonprobabilistic) model takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the number of arcs m as fixed and assumes that all the $\binom{n(n-1)}{m}$ digraphs on V with m arcs are actually possible. Note that $0 \leq m \leq n(n-1)$, and m refers to the total of all out-degrees of the n vertices, and this is also the same as the total of all in-degrees.

Under the present model, the minimum value taken by the out-degree d_i (as well as the in-degree e_i) of the ith vertex is $\max(0, m - (n^2 - 2n + 1))$. To see this, it is enough to notice that the number of arcs left in the complete symmetric digraph, when all the arcs leaving v_i are dropped, is $n^2 - 2n + 1$. Trivially, the maximum value taken by the out-degree d_i (as well as e_i) is $\min(n-1, m)$, and every integer value between the minimum and the maximum can actually be attained by d_i (as well as by e_i).

We now show that the range of $d_{\max}$ is $\{\lceil \frac{m}{n} \rceil, \lceil \frac{m}{n} \rceil + 1, \ldots, \min(n-1, m)\}$. Here $\lceil a \rceil$ denotes the smallest integer greater than or equal to the number a. For example, $\lceil 2 \rceil = 2$ and $\lceil 2.1 \rceil = 3$. That the maximum value of $d_{\max}$ is $\min(n-1, m)$ is trivial to prove.

That $d_{\max}$ in any digraph on n vertices with m arcs is $\geq \lceil \frac{m}{n} \rceil$ follows from $\sum_{i=1}^{n} d_i = m$ since $\max(d_1, d_2, \ldots, d_n) \geq \sum d_i / n$. To construct a digraph G with $d_{\max} = \lceil \frac{m}{n} \rceil$, let $m = nq - r$, where $0 \leq r \leq n-1$. We consider the cases n odd and n even separately. First suppose $n = 2k+1$. Then we show how to partition the set A of arcs of the complete symmetric digraph on $V = \{v_0, v_1, \ldots, v_{2k}\}$ into $n-1$ disjoint subsets of size n, each forming a single cycle (such a cycle is called a *Hamiltonian cycle*). Arrange the vertices $\{v_1, v_2, \ldots, v_{2k}\}$ regularly on a circle with center v_0. The first Hamiltonian cycle is $[v_0, v_1, v_2, v_{2k}, v_3, v_{2k-1}, v_4, \ldots, v_k, v_{k+2}, v_{k+1}, v_0]$. This is shown in the first diagram in Figure 3.2. For $i = 2, 3, \ldots, 2k$, the ith Hamiltonian cycle—namely, $[v_0, v_i, v_{i+1}, v_{i-1}, v_{i+2}, v_{i-2}, v_{i+3}, \ldots, v_{k+i-1}, v_{k+i}, v_0]$—is obtained by rotating the first cycle clockwise around v_0 by $i-1$ steps. Now, we get G by taking the union of q of the Hamiltonian cycles and dropping r arcs of one cycle if $r > 0$. Next let $n = 2k+2$. Then we replace the arc bypassing the central vertex v_0 by a path of length 2 with v_{2k+1} as the middle vertex, as shown in the second diagram in Figure 3.2 (imagine v_{2k+1} to be directly above v_0 in a different plane). As before, we get $2k$ Hamiltonian cycles here also. Now we also have another set of n arcs: the $2k$ arcs in the earlier digraph, which were replaced by paths of length 2 and the two arcs v_0v_{2k+1} and $v_{2k+1}v_0$. Note that every vertex has out-degree 1 w.r.t. the arcs in this last set also. Now, again, we get G by taking the union of q of the Hamiltonian cycles (including the last set if necessary) and dropping r arcs from one set if $r > 0$.

To give the minimum and maximum values taken by the number of sources, let p denote the number of sources. Then min $p = \max(0, n-m)$. To see this, we note that p can be made 0 whenever $m \geq n$ by including a circuit on n vertices in the digraph. If $m \leq n-1$, the m arcs can make at most m vertices nonsources, so $p \geq n-m$, and a digraph with $p = n-m$ is obtained by taking a directed path on $m+1$ vertices together with $n-m-1$ isolated vertices. We now show that max $p = n - \lceil m/(n-1) \rceil$. Denote the right-hand side (RHS) by k. If S is the set of sources in any digraph on n vertices with m arcs, then there is no arc entering any vertex in S, so $m \leq p(n-p) + (n-p)(n-p-1) = (n-1)(n-p)$, which gives $p \leq k$. A digraph G attaining $p = k$ is obtained as follows: Let H be a complete symmetric digraph with vertex set V, where $|V| = n$, and let S be a subset of V with $|S| = k$. Remove all the arcs within S and all the arcs from $V - S$ to S and a few more if necessary (so that exactly m arcs remain) to get G.

Since changing the position of an arc can change the number of sources by at most 1 and since any two digraphs on V with m arcs each can be obtained from each other by changing the position of one arc at a time, it follows that the number p of sources actually takes all integer values between the minimum and the maximum. Finally, it is easy to see that the range of the number of sources is the same as that of the number of sinks.

To give the minimum and maximum values taken by the number of isolates, let q denote the number of isolates. Then it is easy to see that $\min q = \max(0, n-2m)$ since the number of distinct vertices that are end (initial or terminal) vertices of the m arcs is at most $2m$, and the remaining vertices are isolated.We next show that max $q = \ell$, where $\ell = \max\{k : m \leq (n-k)$

$(n-k-1)\}$. Clearly, we have $m \le (n-q)(n-q-1)$ and so $q \le \ell$ for every digraph. To get a digraph attaining ℓ, all one has to do is to put the m arcs within $n-\ell$ vertices. Again, it is easy to see that all values between the minimum and the maximum are attained.

It is easy to see that the minimum and the maximum of the number s of symmetric pairs are $\max\left(0, m-\binom{n}{2}\right)$ and $\lfloor m/2 \rfloor$, respectively, and that every integral value between the minimum and the maximum is attained by s. If t is the number of arcs that are not reciprocated, we have (i) $2s+t=m$ and (ii) $s+t \le \binom{n}{2}$. Subtracting (ii) from (i), we get $s \ge m-\binom{n}{2}$. Since $t \ge 0$, (i) gives $s \le \lfloor m/2 \rfloor$.

The minimum and maximum values of the diameter were obtained by Goldberg and Ghouila-Houri, respectively (see Berge, 1973). If a digraph G is strongly connected, the out-degree as well as the in-degree of every vertex in G is at least 1, so $m \ge n$. Moreover, if $m=n$ in G, then G is a circuit, and the diameter is $n-1$. Hence, the minimum and the maximum values of the diameter are ∞ if $m \le n-1$. The minimum is $n-1$ and the maximum is ∞ if $m=n$. Next we take $n < m \le n(n-1)$. Let $n-1=(m-n+1)q+r$ where $0 \le r < m-n+1$. Then the minimum value of the diameter is

$$\begin{cases} 2q & \text{if } r=0 \\ 2q+1 & \text{if } r=1 \\ 2q+2 & \text{otherwise} \end{cases}, \tag{3.10}$$

and the maximum value of the diameter is

$$\begin{cases} n-1 & \text{if } m \le (n^2+n-2)/2 \\ \left\lfloor n+\frac{1}{2}-\sqrt{2m-n^2-n+\frac{17}{4}} \right\rfloor & \text{otherwise} \end{cases}. \tag{3.11}$$

It is easy to see that the minimum and maximum values of the radius are ∞ if $m < n-1$. If $m \ge n-1$, the minimum radius is 1. For maximum radius, see Berge (1973).

We now prove that the minimum number θ of strong components is

$$\theta = \begin{cases} n & \text{if } m=0 \\ 1 & \text{if } m \ge n \\ n-m+1 & \text{otherwise} \end{cases}. \tag{3.12}$$

The result is trivial if $m \le 1$. If $m \ge n$, then we can get a strongly connected digraph on n vertices with m arcs by starting with a circuit on n vertices and adding $m-n$ arcs arbitrarily, so $\theta=1$. So let $1<m<n$. We first note that any strongly connected digraph on $t \geq 2$ vertices has at least t arcs since the out-degree of each vertex must be at least 1. Let G be any digraph on n vertices with m arcs. Suppose exactly ℓ of the p strong components of G are singletons. If $\ell=n$, then $p=n \geq n-m+1$. If $\ell<n$, then $m \geq n-\ell$, so $p \geq \ell+1 \geq n-m+1$. To get a digraph with exactly $n-m+1$ strong components, consider a circuit on m vertices together with $n-m$ isolated vertices.

We next prove that the maximum number Θ of strong components is

$$k_0=\max\left\{k : k \leq n \text{ and } m-\binom{n}{2} \leq \binom{n-k+1}{2}\right\},$$

which reduces to

$$\left\lfloor \frac{1}{2}\left(2n+1-\sqrt{8\left(m-\frac{n(n-1)}{2}\right)+1}\right)\right\rfloor \text{ if } m \geq \binom{n}{2}. \tag{3.13}$$

(Note that $\binom{1}{2}=0$ by definition.) For this, suppose G is an arbitrary digraph on n vertices with m arcs. Let the strong components of G be $C_1, C_2, \ldots, C_p$, C_i containing n_i vertices. Then no arc between two C_is can be reciprocated, so the number of arcs between C_is is at most $\binom{n}{2}$. Hence,

$$m \leq \binom{n}{2}+\sum_{i=1}^{p}\binom{n_i}{2} \leq \binom{n}{2}+\binom{n-p+1}{2},$$

where the second inequality follows on observing that $\sum_{i=1}^{p}\binom{n_i}{2}$ is maximum when all but one of the n_is are 1 each. So $p \leq k_0$, and it follows that $\Theta \leq k_0$. To show equality, consider the digraph on the vertex set $\{v_1, v_2, \ldots, v_n\}$ obtained as follows: If $m \leq \binom{n}{2}$, put m arcs of the type $v_i v_j$ with $i<j$. If $m>\binom{n}{2}$, make $v_i v_j$ an arc whenever $1 \leq i<j \leq n$ and add $m-\binom{n}{2}$ arcs of the type $v_j v_i$ with $i<j$ within the first $n-k_0+1$ vertices. In this digraph, the last k_0-1 vertices form singleton strong components, so the digraph has at least k_0 and so exactly k_0 strong components.

We now prove that the minimum number ξ of weak components is $\max(1, n-m)$. The result is trivial if $m=0$. If $m \geq n-1$, then we can get a weakly connected digraph on n vertices with m arcs by starting with a path on n vertices and adding $m-n+1$ arcs arbitrarily, so $\xi=1$. Next let $0<m<n-1$. We first note that any weakly connected digraph on t vertices has at least $t-1$ arcs (we omit the proof of this statement). Now let G be any digraph on n

vertices with m arcs. Suppose $C_1, C_2, \ldots, C_q$ are the weak components of G, with C_i having n_i vertices. Then $m \geq \sum_{i=1}^{q}(n_i - 1) = n - q$, so $q \geq n - m$. To get a digraph with exactly $n - m$ weak components, consider a path on $m + 1$ vertices together with $n - m - 1$ isolated vertices.

We next prove that the maximum number Ξ of weak components is

$$\begin{aligned} k^0 &= \max\{k : k \leq n \text{ and } m \leq (n-k+1)(n-k)\} \\ &= \left\lfloor \tfrac{1}{2}\left(2n + 1 - \sqrt{4m+1}\right) \right\rfloor . \end{aligned} \tag{3.14}$$

The proof is similar to that for Θ. Suppose G is an arbitrary digraph with weak components $C_1, C_2, \ldots, C_q$, with C_i containing n_i vertices. Then there are no arcs between C_is, so

$$m \leq \sum_{i=1}^{q} n_i(n_i - 1) \leq (n - q + 1)(n - q),$$

where the second inequality follows as before. Hence, $q \leq k^0$ and $\Xi \leq k^0$. To show equality, consider a digraph with all the arcs belonging to one weak component with $n - k + 1$ vertices.

We now consider the difference $p - q$, where p and q denote the number of strong components and the number of weak components, respectively. We first note that $p \geq q$. If $m \geq 2$, then a digraph with a circuit on $\min(m, n)$ vertices has $p = q$. If $m = 1$, then $p = n$ and $q = n - 1$. If $m = 0$, then $p = q = n$. Hence, we have

$$\min\ p - q = \begin{cases} 1 & \text{if } m = 1 \\ 0 & \text{otherwise} \end{cases} . \tag{3.15}$$

We next show that

$$\max\ p - q = \begin{cases} m & \text{if } m \leq n - 1 \\ \Theta - 1 & \text{otherwise} \end{cases} , \tag{3.16}$$

where Θ is the maximum number of strong components. For this, it is enough to observe that there is a digraph attaining the maximum number of strong components and the minimum number of weak components simultaneously.

We next consider the number h of arcs joining different strong components. The minimum value of h is 0 or 1 accordingly as $m \geq 2$ or $m = 1$. To see this, all one has to do is to include a circuit on $\min(m, n)$ vertices if $m \geq 2$. We next consider the maximum value of h. Clearly, $\max h = m$ if $m \le \binom{n}{2}$. So let $m > \binom{n}{2}$. Let

$$t = \min\left\{ \textstyle\sum_{i=1}^{\ell} \binom{n_i}{2} : \quad \ell \ge 1,\ n_i \ge 1 \text{ for } i = 1, 2, \ldots, \ell, \right.$$
$$\left. \textstyle\sum_{i=1}^{\ell} n_i = n \text{ and } m \le \binom{n}{2} + \sum_{i=1}^{\ell} \binom{n_i}{2} \right\}.$$

Then we will show that $\max h = \binom{n}{2} - t$. Suppose G is a network on n vertices with m arcs, of which h is between strong components. Suppose G has ℓ strong components $C_1, C_2, \ldots, C_\ell$, with C_i containing n_i vertices. Then, for $i \ne j$, there cannot be arcs both from C_i to C_j and from C_j to C_i. Hence, $h \le \binom{n}{2} - \sum_{i=1}^{\ell} \binom{n_i}{2} \le \binom{n}{2} - t$. To prove that the bound is attained, consider a digraph with $v_i v_j$ an arc whenever $i < j$ and with the sizes of the strong components equal to $n_1, n_2, \ldots, n_\ell$ (it is easy to see that such a digraph exists). We mention that $t \ge m - \binom{n}{2}$, but the equality may not always hold. For example, if $n = 6$ and $m = 20$, then $t = 6$ attains when $\ell = 2$ and $n_1 = n_2 = 3$, whereas $m - \binom{n}{2} = 5$. Thus, $\max h$ is 9 here.

We finally come to clique number. It is easy to see that the maximum value for the clique number of a digraph on n vertices with m arcs is

$$\max\{k : m \ge k(k-1)\} = \left\lfloor \tfrac{1}{2}\left(1 + \sqrt{4m+1}\right) \right\rfloor. \tag{3.17}$$

It can be shown that the minimum value of the clique number is

$$\max\left\{k : m \le n(n-1) - k\binom{q}{2} - rq \text{ where } q = \left\lfloor \frac{n}{k} \right\rfloor \text{ and } r = n - kq \right\}. \tag{3.18}$$

We omit the proof of this result. We mention that the corresponding result for (undirected) graphs is known as Turan's theorem. A digraph attaining the bound, which we denote by k, is obtained as follows: The vertex set is $V_1 \cup V_2 \cup \ldots \cup V_k$, where the V_is are pairwise disjoint, r of the V_is has size $q + 1$, and the rest have size q. The m arcs are adjusted in such a way that there is no symmetric pair within any V_i. The clique number of this digraph is k because any clique in it can contain only one vertex from each V_i.

Model II.2

This model is the probabilistic version of Model II.1. It takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the number of arcs m $(0 \le m \le n(n-1))$ as fixed and assumes that all the $\binom{n(n-1)}{m}$ possible digraphs are equally likely. Note that a random digraph is now obtained by choosing m of the $n(n-1)$ ordered pairs $v_i v_j$, by simple random sampling without replacement, and making them arcs.

Since $n(n-1)$ occurs repeatedly, we will write M for $n(n-1)$ in what follows.

Let X_{ij} be defined as before. Then under the present model, $P(X_{ij}=1)=m/M$, but the X_{ij}s are not independent.

Since the out-degree d_i of v_i is the number of pairs chosen from $\{v_i v_j \; 1 \leq j \leq n,\ j \neq i\}$, it follows that d_i has the following hypergeometric distribution:

$$P(d_i = k) = \frac{\binom{n-1}{k}\binom{n^2-2n+1}{m-k}}{\binom{M}{m}},\ \max(0, m-(n^2-2n+1)) \leq k \leq \min(n-1, m). \tag{3.19}$$

Hence, $E(d_i) = m/n$ and

$$V(d_i) = m\frac{1}{n}\left(1-\frac{1}{n}\right)\frac{M-m}{M-1}. \tag{3.20}$$

It may be noted that the distribution of d_i is approximately $B(m, 1/n)$ if $m \leq n-1$ and $B(n-1, m/M)$ if $m \geq n-1$. The distribution can also be approximated by Poisson distribution with mean m/n when $m << n(n-1)$ and a normal distribution with mean and variance given above when m, M, and n are large subject to finite mean and variance in the limits. It is easy to see that the out-degree d_i and the in-degree e_i of v_i have the same distribution.

Now, different d_is are not independent. For example,

$$P(d_2 = k | d_1 = \ell) = \frac{\binom{n-1}{k}\binom{M-2(n-1)}{m-k-\ell}}{\binom{M-(n-1)}{m-\ell}}, \quad k = 0, 1, \ldots, \min(n-1, m-\ell).$$

So the distribution of $d_{\max}$ is not easy to compute now.

The probability that v_i is a source is

$$P(e_i = 0) = \frac{\binom{M-(n-1)}{m}}{\binom{M}{m}}. \tag{3.21}$$

Now, the events that different vertices are sources are not independent. However, the probability that k given vertices are sources is

$$\frac{\binom{(n-k)(n-1)}{m}}{\binom{M}{m}}. \tag{3.22}$$

So an expression similar to (3.1), for the probability that there are exactly k sources, can be written down. A simpler approximation to the distribution of the expected number of sources can be obtained as follows when n is large: We assume that the events that different vertices are sources are independent. Then the number of sources has the distribution $B(n, p)$, where p is given in (3.3). Now,

$$p = \prod_{i=0}^{n-2}\Big(1 - \frac{m}{M-i}\Big) \approx \Big(1 - \frac{m}{M-2n/3}\Big)^{n-1} \approx \exp\Big(-\frac{m(n-1)}{M-2n/3}\Big).$$

This approximation seems to be good if $m < n$. When $m \geq n$,

$$p = \prod_{i=0}^{m-1}\Big(1 - \frac{n-1}{M-i}\Big) \approx \Big(1 - \frac{n-1}{M-2m/3}\Big)^{m} \approx \exp\Big(-\frac{m(n-1)}{M-2m/3}\Big)$$

seems to be better. In both cases, the mean of the true distribution is quite close to the mean of the binomial distribution, but the variance of the true distribution is somewhat smaller than that of the binomial distribution.

The probability that v_i is a sink equals the probability that v_i is a source, and the number of sinks and the number of sources have the same distribution.

The probability that v_i is an isolated vertex is

$$P(d_i = 0 \text{ and } e_i = 0) = \frac{\binom{M-2(n-1)}{m}}{\binom{M}{m}}, \tag{3.23}$$

which is approximately $\exp(-2m(n-1)/(M-4n/3))$ if $m < 2n$. When $m \geq 2n$, $\exp(-2m(n-1)/(M-2m/3))$ is a better approximation. The events that different vertices are isolates are not independent. The probability that k given vertices are isolates is

$$\frac{\binom{(n-k)(n-k-1)}{m}}{\binom{M}{m}}. \tag{3.24}$$

So an expression similar to (3.1), for the probability that there are exactly k isolates, can be written down. However, again, the number of isolates is approximately $B(n, p)$, where p is given by (3.4).

By definition, any two digraphs with m arcs (in particular, any digraph and its converse) have the same probability—namely, $1/\binom{M}{m}$.

The probability that G is symmetric is $\binom{M/2}{m/2}/\binom{M}{m}$, provided that m is even (and 0 otherwise).

The probability that G is asymmetric is $\binom{M/2}{m}2^m/\binom{M}{m}$.

To find the probability that G is complete, we note that given k distinct unordered pairs $\{i_1, j_1\}, \{i_2, j_2\}, \ldots, \{i_k, j_k\}$, the probability that none of

$v_{i_1}v_{j_1}, v_{j_1}v_{i_1}, v_{i_2}v_{j_2}, v_{j_2}v_{i_2}, \ldots, v_{i_k}v_{j_k}, v_{j_k}v_{i_k}$ is an arc is $\binom{M-2k}{m}/\binom{M}{m}$. Hence, the probability that G is not complete is

$$\binom{n}{2}\frac{\binom{M-2}{m}}{\binom{M}{m}} - \binom{\binom{n}{2}}{2}\frac{\binom{M-4}{m}}{\binom{M}{m}} + \binom{\binom{n}{2}}{3}\frac{\binom{M-6}{m}}{\binom{M}{m}} - \cdots \tag{3.25}$$

It is not difficult to see that

$$P(s(G)=k) = \frac{\binom{\binom{n}{2}}{k}\binom{\binom{n}{2}-k}{m-2k}2^{m-2k}}{\binom{n(n-1)}{m}}. \tag{3.26}$$

From this it follows that

$$\frac{P(s(G)=k+1)}{P(s(G)=k)} = \frac{(m-2k)(m-2k+1)}{4(k+1)(\binom{n}{2}-m+k+1)}.$$

This ratio is greater than 1 or less than 1 accordingly as $k < x$ or $k > x$, where

$$x = \frac{(m+4)(m-1)-2n(n-1)}{2n(n-1)+6}. \tag{3.27}$$

Hence, it follows that the distribution of $s(G)$ is unimodal with mode at $\lceil x \rceil$. (The maximum probability may be attained at most at two consecutive integers.) We can also find the mean and the variance since $s(G)$ is the sum of the $n(n-1)/2$ variables Y_{ij}, where $Y_{ij} = X_{ij}X_{ji}$ for any unordered pair $\{i, j\}$ with $i \neq j$. Now

$$E(Y_{ij}) = P(Y_{ij}=1) = \frac{\binom{M-2}{m-2}}{\binom{M}{m}} = \frac{m(m-1)}{M(M-1)}.$$

Let us denote $m(m-1)/(M(M-1))$ by p for convenience. Then

$$E(s(G)) = \frac{Mp}{2} = \frac{m(m-1)}{2n(n-1)-2}. \tag{3.28}$$

Now, $V(Y_{ij}) = p(1-p)$. If $\{i, j\}$ and $\{k, \ell\}$ are distinct, then

$$\begin{aligned}\operatorname{cov}(Y_{ij}, Y_{k\ell}) &= E(Y_{ij}Y_{k\ell}) - E(Y_{ij})E(Y_{k\ell}) \\ &= \frac{m(m-1)(m-2)(m-3)}{M(M-1)(M-2)(M-3)} - p^2.\end{aligned}$$

Noting that $\binom{n}{2} = M/2$, we get

$$V(s(G)) = \frac{M}{2}p(1-p) + \frac{M}{2}\left(\frac{M}{2}-1\right)p\left(\frac{(m-2)(m-3)}{(M-2)(M-3)} - p\right)$$

$$= E(s(G))\left(1 - E(s(G)) + \frac{(m-2)(m-3)}{2M-6}\right). \qquad (3.29)$$

If n is large ($n > 10$, say), $E(s(G)) \approx m^2/(2n^2)$ and $V(s(G)) \approx E(s(G)) \times (1-m/M)^2$. Writing $\alpha = V(s(G))/E(s(G))$ and $\beta = (1-m/M)^2$, we actually have

$$\alpha - \beta = \frac{(M-m)(3(M-m)(M-1) - mM)}{M^2(M-1)(M-3)},$$

so $|\alpha - \beta| \leq 3/M$. Note that the range of $s(G)$ is $[\max(0, m - M/2), [m/2]]$.

It is now even more difficult than in Model I.2 to deal with probabilities of events depending on the distance between vertices.

For example, to find the probability that G is strongly connected, we have to find the number $g(n, m)$ of strongly connected digraphs on the vertex set $\{v_1, v_2, \ldots, v_n\}$ with m arcs. No method for computing this number $g(n, m)$ is known. Thus, even finding the probability that G is strongly connected is difficult. If $n = 3$ and $m \leq 2$, the probability is 0. If $n = 3$ and $m = 3$, out of the 20 possible digraphs, only 2 are strongly connected, so the probability is $1/10$. If $n = 3$ and $m = 4$, out of the 15 possible digraphs, 9 are strongly connected, so the probability is $3/5$. If $n = 3$ and $m \geq 5$, then the probability is 1.

The probability that G has diameter 1 is clearly 1 or 0 according to whether $m = n(n-1)$ or not. When $n = 3$ and $m = 5$, G has diameter 2 with probability 1. When $n = 3$ and $m = 4$, G has diameter 2 with probability $3/5$ and diameter ∞ with probability $2/5$. When $n = 3$ and $m = 3$, G has diameter 2 with probability $1/10$ and diameter ∞ with probability $9/10$. When $n = 3$ and $m \leq 2$, G has diameter ∞ with probability 1.

When $n = 3$ and $m \geq 4$, G has radius 1 with probability 1. When $n = 3$ and $m = 3$, G has radius 1 with probability $3/10$ and radius 2 with probability $7/10$. When $n = 3$ and $m = 2$, G has radius 1 with probability $1/5$, radius 2 with probability $2/5$, and radius ∞ with probability $2/5$. When $n = 3$ and $m \leq 1$, G has radius ∞ with probability 1.

Results of Simulation

To give an idea of the distributions of various statistics when both n and m are fixed, we give them for $n = 10$ and a few values of m. These were obtained by simulation using 100,000 random digraphs.

Maximum Out-Degree (Continuation)

	Maximum Out-Degree								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
m									
4	.538	.434	.027	.001	–	–	–	–	–
7	.076	.724	.184	.015	.001	–	–	–	–
10	.001	.478	.444	.071	.006	–	–	–	–
15	–	.055	.583	.305	.052	.005	–	–	–
20	–	–	.228	.546	.193	.030	.003	–	–
40	–	–	–	–	.122	.529	.291	.054	.004
60	–	–	–	–	–	–	.172	.630	.198
80	–	–	–	–	–	–	–	.001	.999

Sources

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
m										
4	–	–	–	–	–	–	.538	.411	.050	.001
7	–	–	–	.076	.359	.412	.140	.013	–	–
10	.001	.024	.178	.392	.309	.088	.008	–	–	–
15	.075	.319	.394	.180	.030	.002	–	–	–	–
20	.323	.461	.189	.026	.001	–	–	–	–	–
40	.964	.036	–	–	–	–	–	–	–	–
60	1	–	–	–	–	–	–	–	–	–
80	1	–	–	–	–	–	–	–	–	–

Isolates

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*
m								
4	–	–	.029	.236	.450	.248	.036	.001
7	.046	.262	.414	.231	.044	.003	–	–
10	.312	.465	.195	.027	.001	–	–	–
15	.764	.221	.015	–	–	–	–	–
20	.940	.059	.001	–	–	–	–	–
40	1	–	–	–	–	–	–	–
60	1	–	–	–	–	–	–	–
80	1	–	–	–	–	–	–	–

	Symmetric Pairs							
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*
m								
4	.932	.067	.001					
7	.775	.212	.013	–	–	–	–	–
10	.571	.356	.069	.004	–	–	–	–
15	.247	.419	.254	.070	.009	.001	–	–
20	.065	.241	.337	.240	.094	.021	.002	–

	Symmetric Pairs (Continued)										
	4	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
$m = 40$	.004	.019	.062	.137	.213	.238	.181	.098	.037	.009	.002

	Symmetric Pairs (Continued)									
	16	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*
$m = 60$	.008	.041	.126	.227	.261	.200	.097	.033	.006	.001

	Symmetric Pairs (Continued)			
	35	*36*	*37*	*38*
$m = 80$	.573	.356	.067	.004

	Diameter									
	1	2	3	4	5	6	7	8	9	∞
m										
4	–	–	–	–	–	–	–	–	–	1
7	–	–	–	–	–	–	–	–	–	1
10	–	–	–	–	–	–	–	–	–	1
15	–	–	–	–	–	.001	.001	.001	–	.997
20	–	–	–	.002	.028	.032	.015	.004	.001	.918
40	–	.001	.738	.181	.009	–	–	–	–	.071
60	–	.880	.119	–	–	–	–	–	–	.001
80	–	1	–	–	–	–	–	–	–	–

	Radius								
	1	2	3	4	5	6	7	8	∞
m									
4	–	–	–	–	–	–	–	–	1
7	–	–	–	–	–	–	–	–	1
10	–	–	.001	.002	.002	.002	.001	–	.992
15	–	.008	.092	.093	.054	.023	.007	.002	.721
20	–	.105	.389	.149	.057	.016	.004	.001	.279
40	.004	.987	.009	–	–	–	–	–	–
60	.199	.801	–	–	–	–	–	–	–
80	.999	.001	–	–	–	–	–	–	–

	Clique Number							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
m								
4	.932	.068	–	–	–	–	–	–
7	.775	.225	–	–	–	–	–	–
10	.571	.429	–	–	–	–	–	–
15	.247	.752	.001	–	–	–	–	–
20	.065	.928	.007	–	–	—	–	–
40	–	.492	.504	.004	–	–	–	–
60	–	–	.437	.544	.019	–	–	–
80	–	–	–	.001	.253	.645	.100	.001

	Strong Components									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
m										
4	–	–	–	–	–	–	–	.009	.067	.924
7	–	–	–	–	–	.004	.021	.074	.195	.706
10	–	–	–	.004	.018	.051	.104	.177	.251	.395
15	.003	.022	.069	.127	.166	.170	.158	.130	.095	.060
20	.082	.209	.250	.196	.126	.070	.037	.019	.008	.003
40	.929	.068	.003	–	–	–	–	–	–	–
60	1	–	–	–	–	–	–	–	–	–
80	1	–	–	–	–	–	–	–	–	–

	Weak Components									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
m										
4	–	–	–	–	–	.897	.102	.001	–	–
7	–	–	.444	.460	.092	.004	–	–	–	–
10	.197	.504	.262	.036	.001	–	–	–	–	–
15	.750	.234	.016	–	–	–	–	–	–	–
20	.939	.060	.001	–	–	–	–	–	–	–
40	1	–	–	–	–	–	–	–	–	–
60	1	–	–	–	–	–	–	–	—	–
80	1	–	–	- -	–	–	–	–	–	–

3.4 MODELS FIXING ALL OUT-DEGREES OF INDIVIDUAL VERTICES

Model III.1

Model II.1 leaves the possibility that all the m arcs are within a few vertices. This is often not realistic, and many of the vertices may have positive out-degree. In such a situation, as well as others in which the respondent is asked a question to name his or her three best friends, Model III.1 is more appropriate.

This model takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the out-degree d_i of v_i as fixed for $i = 1, 2, \ldots, n$ and assumes that all the $\prod \binom{n-1}{d_i}$ digraphs on V with $d^+(v_i) = d_i$ for $i = 1, 2, \ldots, n$ are actually possible. Note that the d_is have to satisfy the following condition: $0 \le d_i \le n - 1$ for all i. For the sake of convenience, we will assume in the discussion of this model that $d_1 \ge d_2 \ge \cdots \ge d_n$. Moreover, G will denote a digraph with vertex set $\{v_1, v_2, \ldots, v_n\}$ and with $d^+(v_i) = d_i$ for all i.

Clearly, the present model fixes m since $m = \sum_{i=1}^{n} d_i$. The number of sinks is also fixed since v_i is a sink if and only if $d_i = 0$. We now show that the range of the in-degree e_j of v_j is $[\sum_{i \ne j} \max(0, d_i - n + 2), \sum_{i \ne j} \min(1, d_i)]$. Clearly, $e_j \ge \sum_{i \ne j} \max(0, d_i - n + 2)$ in every G since for all $i \ne j$ such that $d_i > n - 2$, $v_i v_j$ is an arc. It is easy to see that the bound is attained. Also, $e_j \le \sum_{i \ne j} \min(1, d_i)$ in every G since for all $i \ne j$, $v_i v_j$ can be an arc only if $d_i \ge 1$. Again it is easy to see that the bound is attained.

To give the minimum and maximum values taken by the number of sources, let p denote the number of sources. Then we show that min $p = \max(0, n - m)$. Suppose G has a source u and a vertex v with in-degree at least 2. We may then take an arc xv where $x \neq u$, drop the arc xv, and introduce the arc xu. This will give another digraph with the same out-degrees and with less number of sources. Now suppose $m \geq n$. If there is a source in G, then some vertex has in-degree at least 2, so we can reduce the number of sources. Repeating the process, we get a digraph with $p = 0$. Next let $m < n$. Then the m arcs can make at most m vertices nonsources, so $p \geq n - m$ in every G. By the argument given above, a digraph with the minimum number of sources has no vertex with in-degree larger than 2 and so has $n - m$ sources. Hence, min $p = n - m$.

To give the maximum value taken by p, let $k = d_1$. Then max $p = n - k$ if $d_{n-k+1} < k$ and $n - k - 1$ otherwise. To see this, we note that in any G, v_1 is joined to k vertices, so $p \leq n - k$. If $d_{n-k+1} < k$, then v_i can be joined to d_i of the last k vertices for $i = 1, 2, \ldots, n$, giving a digraph with $p = n - k$. Next let $d_{n-k+1} = k$. Then in any G, if W is the set of vertices to which v_1 is joined, the out-degree of at least one vertex in W is larger than $k - 1$, so at least one vertex outside W is not a source, and hence $p \leq n - k - 1$. It is easy to see that this bound is attained.

To give the minimum and maximum values taken by the number of isolates, let q denote the number of isolates. Let $d_1 = k$ and let ℓ be the number of d_is, which are 0. We show that $\min q = \max(0, \ell - \sum_{i=1}^{n-\ell} d_i)$. In any G, at most $\sum_{i=1}^{n-\ell} d_i$ of the last ℓ vertices can have positive in-degree, and hence it follows that $q \geq \max(0, \ell - \sum_{i=1}^{n-\ell} d_i)$. A digraph attaining the bound is obtained by choosing a vertex among the last ℓ, which has not yet received any arc (if it exists) while choosing the vertices to which v_i is joined, $i = 1, 2, \ldots, n - \ell$.

We now give the maximum value taken by the number of isolates. If $k = 0$, then $\max q = n$. So let $k \geq 1$. Then we show that $\max q = \min(n - k - 1, \ell)$. Clearly, v_1 and the vertices to which it is joined as well as the $n - \ell$ vertices with positive out-degree are not isolates, so $q \leq \min(n - k - 1, \ell)$. Clearly, the bound is attained by the digraph obtained by joining v_i to d_i of the first $k + 1$ vertices for $i = 1, 2, \ldots, n - \ell$. It is easy to see that every value between the minimum and the maximum values of q is attained.

The range of $s(G)$ in the present model was determined by Achuthan, Rao, and Rao (1984). To give this range, let the d_is be arranged so that $d_1 \geq d_2 \geq \cdots \geq d_n$. For any t with $1 \leq t \leq n$, define

$$f(t) = \sum_{i=1}^{t} d_i - t(n - t) - \binom{t}{2}. \tag{3.30}$$

Then the minimum value of the number $s(G)$ of reciprocal pairs is $\max_t f(t)$. We show only that $\min s(G) \geq \max_t f(t)$. Fix any t and any digraph with vertex set $\{v_1, v_2, \ldots, v_n\}$ such that the out-degree of v_i is d_i for $i = 1, 2, \ldots, n$. Then the number of arcs $v_i v_j$ in G with $i \leq t$ is $\sum_{i=1}^{t} d_i$. Of these, at most $t(n-t)$ can have $j \geq t+1$. So G has at least $\sum_{i=1}^{t} d_i - t(n-t)$ arcs within the first t vertices. Hence, G has at least $f(t)$ symmetric pairs (within the first t vertices). This proves $\min s(G) \geq \max_t f(t)$.

To give the maximum value of $s(G)$, define

$$g(t) = \sum_{i=1}^{t} d_i - t(t-1) - \sum_{j=t+1}^{n} \min(t, d_j). \tag{3.31}$$

Then, $\max s(G) = (m - \max_t g(t))/2$, where $m = \sum_{i=1}^{n} d_i$. We only show that the left-hand side does not exceed the right-hand side. Fix t and G as in the preceding paragraph. Then the number of arcs $v_i v_j$ in G with $i \leq t$ is $\sum_{i=1}^{t} d_i$. Of these, at most $t(t-1)$ can have $j \leq t$. So G has at least $\sum_{i=1}^{t} d_i - t(t-1)$ arcs from the first t vertices to the last $n-t$ vertices. Now there are at most $\sum_{j=t+1}^{n} \min(t, d_j)$ arcs from the last $n-t$ vertices to the first t vertices. Hence, there are at least $g(t)$ arcs from the first t vertices to the last $n-t$ vertices, which are not reciprocated (i.e., do not form symmetric pairs). Hence, $s(G) \leq (m - g(t))/2$.

It is easy to see that every value between the minimum and the maximum values of $s(G)$ is attained.

Exact analytical determination of the range of the diameter seems to be a difficult problem when the out-degrees of the vertices are fixed.

We now give the minimum value ρ of the radius r. If $\sum_i d_i < n-1$, clearly $\rho = \infty$. So let $\sum_i d_i \geq n-1$. Let u be a center of G. Clearly, only $n_0 = 1$ vertex is at distance 0 from u. It is easy to see that at most, $n_1 := n_0 + d_1$ vertices are at distance ≤ 1 from u, at most $n_2 := n_1 + d_{n_0+1} + d_{n_0+2} + \cdots + d_{n_1}$ vertices are at distance ≤ 2 from u, at most $n_3 := n_2 + d_{n_1+1} + d_{n_1+2} + \cdots + d_{n_2}$ vertices are at distance ≤ 3 from u, and so on. It follows easily that ρ is the smallest k such that $n_k \geq n$. To give an example, suppose the out-degrees are 3,2,2,1,1,1,1,1,1,0. Then, $n_0 = 1$, $n_1 = 4$, $n_2 = 9$, and $n_3 = 14$. Since $n = 10$, it follows that $\rho = 3$. We do not know the maximum value of the radius.

We now prove that the minimum number of strong components is min $(n, \ell + 1)$, where ℓ is the number of d_is, which are 0. Clearly, any G will have at least $\min(n, \ell + 1)$ strong components. We can get a digraph attaining this bound by including a circuit on the first $n - \ell$ vertices if $n - \ell \geq 2$.

We do not have a simple expression for the maximum number of strong components, but we will show that the digraph D with the property

$P : v_j$ is joined *to* the last d_j v_is (excluding v_j itself) for $j = 1, 2, \ldots, n$,

is extremal. We prove this by induction on n. The basis for induction is trivial. Assume the result for less than n vertices. Let G be any extremal digraph (i.e., a digraph with $d^+(v) = d_i$ for $i = 1, 2, \ldots, n$ and with the maximum number k of strong components). Then the strong components $C_1, C_2, \ldots, C_k$ in G can be numbered in such a way that (i) there is no arc from C_i to C_j if $i > j$. To see this, we note that the condensation H of G obtained by shrinking each C_i to a vertex has no circuits. Let P be an open (i.e., the first and last vertices are distinct) path of the maximum length in H. Then the first vertex of P is a source in H. Let C_1 be the strong component of G corresponding to this vertex. Now deleting C_1 and repeating the procedure, we can order the strong components so that (i) is satisfied. Now, if $u \in C_i$, $v \in C_j$, $i < j$, and $d^+(u) < d^+(v)$, then by interchanging u and v (note that condition (i) is easy to satisfy), we get another extremal digraph. Repeating this process, we get an extremal digraph G such that the last few (say, n_k) v_is form a strong component. Clearly, we may assume that each v_j belonging to C_k is joined *to* the last d_j v_is (excluding v_j itself). We may also assume that for all $i \leq n - n_k$, v_i is joined *to* the last $\min(d_i, n_k)$ v_is. Finally, the subdigraph induced by $C_1 \cup \cdots \cup C_{k-1}$ is also extremal and can be replaced by one satisfying P. Then G itself satisfies P. The number of strong components in such a digraph can easily be counted, although it is difficult to give a nice expression for it. For example, suppose the out-degree sequence is (9, 8, 8, 5, 5, 5, 5, 3, 2, 0). Then the strong components in the extremal digraph obtained as above are $\{v_{10}\}$, $\{v_5, v_6, v_7, v_8, v_9\}$, $\{v_4\}$, $\{v_2, v_3\}$, and $\{v_1\}$ (note that it is convenient to determine the strong components starting at the right end).

Let p denote the number of weak components. Then we show that $\min p = \max(1, n - m)$. Suppose G has two weak components, C_1 and C_2, and C_1 is not a tree. Then C_1 has an arc uv such that C_1 remains weakly connected when arc uv is deleted. Now deleting the arc uv and joining u to a vertex of C_2 reduces the number of weak components. Thus, in a digraph attaining the minimum number of weak components, either there is only one weak component or there are at least two weak components, and all of them are trees. In the former case, $m \geq n - 1$, and in the latter case, $m = n - p \leq n - 2$. This proves that $\min p = \max(1, n - m)$.

We do not know the maximum value Ξ of the number of weak components precisely but have partial information on it. Let G be a maximal digraph (i.e., a digraph attaining Ξ). Let C_1 be the weak component containing v_1. If there is a vertex u other than v_1 in C_1 and a vertex v in another weak component such that the out-degree of u is less than the out-degree of v, then by interchanging u and v (note that this is possible since the out-degree of v is less than the size of C_1), we get another extremal digraph. Repeating this process, we get an extremal digraph such that the vertices in C_1 have the largest few out-degrees. Repeating the argument starting with the first vertex not in C_1, we ultimately get an extremal digraph in which the weak components are $\{1, 2, \ldots, n_1\}$, $\{n_1+1, n_1+2, \ldots, n_2\}, \ldots, \{n_{p-1}+1, n_{p-1}+2, \ldots, n\}$ for some $n_1, n_2, \ldots, n_p$ such that $1 \leq n_1 < n_2 < \cdots < n_p = n$ and $n_{i+1} - n_i \geq d_{n_i+1} + 1$ for $i = 0, 2, \ldots, p-1$. It is also easy to see that Ξ is the maximum p such that such n_is exist. However, we cannot always take $n_1 = d_1 + 1$ and so forth. For example, if the d_is are 3, 3, 3, 2, 2, 2, we have to take $n_1 = 6$ and $\Xi = 1$. Moreover, if the d_is are 3, 3, 3, 3, 3, 1, 1, 0, taking $n_1 = 5, n_2 = 7$, and $n_3 = 8$ is better than taking $n_1 = 4$ and $n_2 = 8$.

We finally come to clique number ω. We show that $\max \omega = \max\{k : d_k \geq k-1\}$. For this, it is enough to observe that k vertices can be made to induce a complete symmetric digraph if and only if their out-degrees are at least $k-1$. The minimum value of ø is not known.

Model III.2

This model takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the out-degree d_i $(0 \leq d_i \leq n-1)$ of v_i as fixed for $i = 1, 2, \ldots, n$ and assumes that all the $\prod \binom{n-1}{d_i}$ possible digraphs are equally likely.

Note that now a random digraph is obtained by choosing, for each i, d_i of the $n-1$ ordered pairs $v_i v_j$, $1 \leq j \leq n$, $j \neq i$, by simple random sampling without replacement and making them arcs. Note that different is are treated independently.

Let X_{ij} be defined as before. Then, under the present model, $P(X_{ij} = 1) = d_i/(n-1)$. Also, X_{ij} and $X_{k\ell}$ are independent if $i \neq k$.

Clearly, now d_is and thus m are fixed. But the in-degrees are variable, and the distribution of e_j is not easy to compute. However, the mean and the variance of e_j can be computed as follows Since $e_j = \sum\{X_{ij} : 1 \leq i \leq n, i \neq j\}$, it follows that $E(e_j) = \sum_{i \neq j} E(X_{ij}) = \sum_{i \neq j} d_i/(n-1) = (m - d_j)/(n-1)$. Also, since X_{ij}s are independent for different is, we get

$$V(e_j) = \sum_{i \neq j} V(X_{ij}) = \sum_{i \neq j} \frac{d_i(n-1-d_i)}{(n-1)^2} = \frac{(n-1)(S_1 - d_j) - (S_2 - d_j^2)}{(n-1)^2}, \tag{3.32}$$

where $S_1 := \sum d_i = m$ and $S_2 := \sum d_i^2$.

The probability that v_j is a source is

$$P(e_j = 0) = \prod_{i\,:\,i \neq j} \left(1 - \frac{d_i}{n-1}\right). \tag{3.33}$$

Even though one can, in principle, write an expression for any k-given vertices to be sources, this and the expression obtained from it for the probability that there are exactly k sources are not useful. Note that the e_js are not independent since their sum is a constant. So the distribution of $e_{\max}$ is also not easy to compute.

The probability that v_j is an isolated vertex is $P(e_i = 0)$ if $d_i = 0$ and 0 otherwise. The events that different vertices are isolates are not independent. The probability that k-given vertices are isolates and the probability that there are exactly k isolates are not easy to compute.

Now the probabilities that G is symmetric, asymmetric, complete, and so on are all difficult to find.

For any unordered pair $\{i, j\}$, the probability that none of $v_i v_j$ and $v_j v_i$ is an arc is

$$\frac{(n-1-d_i)(n-1-d_j)}{(n-1)^2}.$$

But given k distinct unordered pairs $i_1 j_1, i_2 j_2, \ldots, i_k j_k$, the probability that none of $v_{i_1} v_{j_1}, v_{j_1} v_{i_1}, v_{i_2} v_{j_2}, v_{j_2} v_{i_2}, \ldots, v_{i_k} v_{j_k}, v_{j_k} v_{i_k}$ is an arc is not easy to write down. Hence, the probability that G is not complete cannot be found easily.

The distribution of the number of symmetric pairs $s(G)$ is complicated, but its mean and variance were computed by Katz and Wilson (1956). To compute these, we write $s(G)$ as the sum of Y_{ij}s as before. Now $E(Y_{ij}) = d_i d_j/(n-1)^2$. Hence, writing $S_k = \sum d_i^k$ for $k = 1, 2, \ldots$, we get

$$E(s(G)) = \frac{S_1^2 - S_2}{2(n-1)^2}. \tag{3.34}$$

It can be proved similarly that

$$V(s(G)) = E(s(G)) + \frac{S_1^2 S_2 - S_2^2 - 2S_1S_3 + 2S_4 - S_1^3 + 3S_1S_2 - 2S_3}{(n-1)^3(n-2)} - \frac{2S_1^2 S_2 - S_2^2 - 4S_1S_3 + 3S_4}{2(n-1)^4}. \quad (3.35)$$

If the d_is do not differ much, it can be seen that

$$E(s(G)) \approx \frac{n\bar{d}^2}{2(n-1)} \text{ and } V(s(G)) \approx E(s(G))\left(1 - \frac{\bar{d}}{n-1}\right)^2,$$

where $\bar{d}$ denotes $\sum d_i/n$.

Model III.3

This model is a generalization of Model III.2. Here, again, we take the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the out-degree d_i $(0 \le d_i \le n-1)$ of v_i as fixed for $i = 1, 2, \ldots, n$. But we now assume that for each i, there is a set $P_i \subseteq \{1, 2, \ldots, n\} - \{i\}$ and that the terminal vertices of the d_i arcs leaving v_i belong to P_i. Finally, the model assumes that the $\prod \binom{n_i}{d_i}$ possible digraphs are equally likely, where n_i denotes the size of P_i.

The set P_i may be called the *potential set* for v_i since v_i makes its choices from P_i. The need for considering this type of model has already been explained in Chapter 1. Model III.2 implicitly makes two important assumptions: (i) for any fixed i, the probability that v_j is chosen by v_i is the same for all j, and (ii) the choices of different v_is are statistically independent. Assumption (i) is perhaps unrealistic when d_i is much smaller than n. So we remove it in the present model while still retaining assumption (ii).

Let X_{ij} be defined as before. We take X_{ii} to be 0. Then, under the present model, $P(X_{ij} = 1)$ is d_i/n_i if $j \in P_i$ and 0 otherwise. Also, X_{ij} and $X_{k\ell}$ are independent if $i \ne k$.

Clearly, again, d_is and thus m are fixed. But the in-degrees are variable, and the distribution of e_j is not easy to compute. However, the mean and the variance of e_j can be computed as follows. Let us denote d_i/n_i by r_i. Then $E(X_{ij}) = r_i$ and $V(X_{ij}) = r_i(1 - r_i)$ if $j \in P_i$. If $j \notin P_i$, then $E(X_{ij}) = V(X_{ij}) = 0$. Now, $e_j = \sum_i X_{ij}$, so $E(e_j) = \sum_{i\,:\,j \in P_i} r_i$. Also, since X_{ij}s are independent for different is, we get

$$V(e_j) = \sum\{r_i(1 - r_i) : j \in P_i\}. \quad (3.36)$$

Although the distribution of s is difficult to compute, Rao and Bandyopadhyay (1987) showed that $E(s) = \sum_{j=1}^{n} r_j T_j/2$ and

$$V(s) = \frac{1}{2}\sum_{j=1}^{n} r_j T_j - \frac{1}{2}\sum_{j=1}^{n} r_j^2 W_j - \sum_{j=1}^{n} \frac{r_j(1-r_j)}{n_j - 1}(T_j^2 - W_j), \qquad (3.37)$$

where

$$T_j = \sum\{r_i : i \in P_j \text{ and } j \in P_i\} \text{ and } W_j = \sum\{r_i^2 : i \in P_j \text{ and } j \in P_i\}. \qquad (3.38)$$

If we make the assumptions (i) $i \in P_j$ if $j \in P_i$ and (ii) $r_j = r$ for all j, then

$$E(s) = \frac{mr}{2} \text{ and } V(s) = \frac{mr}{2}(1-r)^2.$$

3.5 MODELS FIXING ALL OUT-DEGREES AND IN-DEGREES OF INDIVIDUAL VERTICES

Model IV.1

This model takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the out-degree d_i as well as the in-degree e_i of v_i as fixed for $i = 1, 2, \ldots, n$ and assumes that all the digraphs on V with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$ are actually possible. Note that the d_is and e_is have to satisfy the following conditions: $0 \le d_i \le n-1$ and $0 \le e_i \le n-1$ for all i and $\sum d_i = \sum e_i$. But these conditions are not sufficient. Ryser (1963) and Fulkerson (1966) proved that these together with the following condition are necessary and sufficient for the existence of a digraph with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$:

$$\sum_{i=1}^{k} d_i \le \sum_{j=1}^{k} \min(k-1, e_j) + \sum_{j=k+1}^{n} \min(k, e_j), \quad k = 1, 2, \ldots, n, \qquad (3.39)$$

where we assume without loss of generality that $d_1 \ge d_2 \ge \cdots \ge d_n$. We will make this assumption for convenience in what follows. Moreover, G will denote a digraph with vertex set $\{v_1, v_2, \ldots, v_n\}$ and with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for all i.

We will prove only the necessity of the above condition. Clearly, there are $\sum_{i=1}^{k} d_i$ arcs $v_i v_j$ with $1 \le i \le k$. The number of arcs among these with $1 \le j \le k$ is at most $\sum_{j=1}^{k} \min(k-1, e_j)$, and the number with $k+1 \le j \le n$ is at most $\sum_{j=k+1}^{n} \min(k, e_j)$, and hence the inequality follows. This proves the necessity. A proof of sufficiency can be found in Berge (1973). One can easily give an example where the inequality is not satisfied: Take both the out-degree and in-degree sequences to be (3, 3, 1, 1). Then the inequality is violated for $k = 2$.

Under the present model, the total number of arcs, out-degrees, in-degrees, and sources, sinks, and isolates is all fixed. The range of the other parameters considered earlier seem to be difficult to find. Rao (1984; unpublished manuscript) gave some lower and upper bounds for $\min s(G)$. These are quite involved, and we only mention a special case. Let $a_i = d_i + e_i$ for all i. If the product of the two largest a_is does not exceed $N + 1$, where N is the number of nonzero a_is, then $\min s(G) = 0$. We next note that $\max s(G) \leq \lfloor \sum_{i=1}^{n} \min(d_i, e_i)/2 \rfloor$. To prove this, it is enough to note that there can be at most $\min(d_i, e_i)$ symmetric pairs containing v_i. We do not know any sufficient conditions under which the bound is attained, and we do not have any lower bound for $\max s(G)$. We give an example to show that the above upper bound is not always attained. Suppose there is a digraph G with out-degree and in-degree sequences (3, 2, 2, 2, 1, 1, 0) and (0, 0, 0, 2, 2, 3, 4) and with two symmetric pairs. Then v_4v_5, v_4v_6, v_5v_4, and v_6v_4 should be arcs, and v_7 cannot have in-degree 4, a contradiction. Note, however, that there exists a digraph with the given out-degree and in-degree sequences and with one symmetric pair.

Even though we do not know the minimum and maximum values of $s(G)$ in general, we may be able to find them sometimes by using the technique described below. Suppose we are given one digraph G. By an *alternating rectangle* in G, we mean four distinct vertices u, v, w, x such that uv and wx are arcs and ux and wv are not arcs. (Note that the entries in the cells of the adjacency matrix corresponding to the pairs (u, v), (u, x), (w, x), and (w, v) are 1, 0, 1, 0—hence the name *alternating rectangle*.) By *switching along this alternating rectangle*, we mean dropping the arcs uv and wx and introducing the arcs ux and wv (this amounts to interchanging 1s and 0s in the four cells of the adjacency matrix referred to above). Note that this does not alter the out-degree or the in-degree of any vertex. Similarly, by a *compact alternating hexagon* in G, we mean three distinct vertices u, v and w such that uv, vw, and wu are arcs and vu, wv, and uw are not arcs. (Note that the entries in the cells of the adjacency matrix corresponding to the pairs (u, v), (w, v), (w, u), (v, u), (v, w), and (u, w) are 1, 0, 1, 0, 1, 0—hence the name *compact alternating hexagon*.) By *switching along this compact alternating hexagon*, we mean dropping the arcs uv, vw, and wu and introducing the arcs vu, wv, and uw. Again, this does not alter the out-degree or the in-degree of any vertex. Rao, Jana, and Bandyopadhyay (1996) proved that any digraph H with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for all i can be obtained from any other such digraph G by a finite sequence of switches along alternating rectangles and compact alternating hexagons. (In most cases, this can be achieved by using only alternating rectangles.) Using this technique, we may be able to prove in some cases by actual construction that a known lower bound or upper bound for $s(G)$ is actually attained.

We do not know the range of the diameter, the radius, and the number of strong and weak components when both d_is and e_is are fixed. However, we can give the minimum number of weak components. For this, let $m = \sum d_i = \sum e_i$, and let p denote the number of weak components. Let ℓ be the number of is such that $d_i = e_i = 0$. Then min $p = \ell + \max(1, n - \ell - m)$ (when $m < n - \ell - 1$, every weak component of a digraph attaining min p is a "tree" on one or more vertices; when $m \geq n - \ell - 1$, a digraph attaining min p has only one weak component other than the ℓ isolated vertices), but the value of max p is not known. Finally, the range of the clique number is also not known.

Model IV.2

This model takes the vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and the out-degree d_i as well as the in-degree e_i of v_i as fixed for $i = 1, 2, \ldots, n$ and assumes all the digraphs on V with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$ to be equally likely.

Note that the d_is and e_is have to satisfy the conditions stated in Model IV.1 for the existence of at least one digraph with the given out-degrees and in-degrees. (If the out-degrees and in-degrees are read from an observed network, these conditions will automatically be satisfied.) The number of possible networks is a very complicated function of d_is and e_is and is not known, although some recursive methods have been given by Sukhatme (1938) and Katz and Powell (1954) for computing it. The number can be astronomical even for moderate n like $n = 40$. There is also no easy way of generating (listing) all the networks.

Under the present model, practically nothing is known about the distribution of any of the statistics except those that are trivially fixed by the model. So one has to take recourse to simulation. But even generating a random network (where all possible networks are chosen with equal probabilities) is a difficult problem now.

Pramanik (1994) gave a heuristic procedure to get an approximately random network by generating its incidence matrix. His method basically consists of the following: choose a cell, the (i, j)th with probability proportional to $d_i e_j$; put a 1 in that cell; update the d_is and e_is; and repeat. At each stage including the beginning, the entries in all the cells that are determined by the current d_is and e_is are filled in before applying the above procedure. Although this procedure seems to give an approximately random network, there is no theoretical basis for it or an estimate of how good the approximation is.

Snijders (1991) bypassed the problem of generating a random matrix and gave a way of generating a nearly random network so that its probability could be computed. One can then estimate the distribution of any statistic (under the model where all possible networks are equally likely) using a ratio estimator.

One possible way of generating a random network is the following: Generate a random network with out-degree sequence $(d_1, d_2, \ldots, d_n)$. Accept it if its in-degree sequence is $(e_1, e_2, \ldots, e_n)$; otherwise, reject it, generate another with out-degree sequence $(d_1, d_2, \ldots, d_n)$, and repeat this until a network is accepted. It is easy to see that this gives a random network under the present model, but the rate of rejection will generally be so high that not even one may be accepted in a million even for moderate values of n.

Rao et al. (1996) gave a Markov chain simulation method for generating a random network. We will briefly describe the method without detailed proofs. In the following text, we use the term *alternating cycles* to mean alternating rectangles and compact alternating hexagons as defined in the discussion of Model IV.1.

Let $\mathfrak{G}$ denote the set of all networks with vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and with $d^+(v_i) = d_i$ and $d^-(v_i) = e_i$ for $i = 1, 2, \ldots, n$. The basic step of the Markov chain Monte Carlo (MCMC) method is as follows. We start with an initial network belonging to $\mathfrak{G}$. At any stage, we enumerate the alternating cycles in the current network, choose one of them at random, and switch along it to get a new network in $\mathfrak{G}$. We perform this process a large number of times. Then one believes that the network obtained should be a random network (we shall see presently that this is not quite correct). To try to prove this, let us formulate it as a Markov chain. The states of the Markov chain are the networks belonging to $\mathfrak{G}$. We shall say that two states are adjacent if the networks represented by them can be obtained from each other by switching along one alternating cycle. Let $c(i)$ denote the number of states adjacent to the state i. Then the procedure mentioned above obviously gives a Markov chain with transition probability $p_{ij} = 1/c(i)$ if j is adjacent to i and 0 otherwise. Note that the Markov chain is irreducible since every state can be reached from every other state in a finite number of steps. So, there exists a unique stationary distribution (see Feller, 1968), and if the Markov chain is aperiodic, the distribution of the state after q steps approaches this stationary distribution as $q \to \infty$, whatever the initial state. To find the stationary distribution, note that it is a probability vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_N)$ such that $\pi'\mathbf{P} = \pi'$, where $\mathbf{P}$ is the transition probability matrix $((p_{ij}))$. Taking $\theta_i = c(i)/\sum_k c(k)$, it is easy to see that $\sum_i \theta_i p_{ij} = \theta_j$, and thus $\pi_i = \theta_i$. Thus, according to the stationary distribution, the probability of the ith state is not $1/N$ but is proportional to the number $c(i)$ of alternating cycles in the network corresponding to the ith state.

As observed in the preceding paragraph, the basic Markov chain simulation method given above does not choose the networks in $\mathcal{E}$ with equal probabilities and needs modification. Suppose we know an upper bound K for the $c(i)$s. Then we can modify the basic method as follows: At any stage, if we are currently in state i, we go to any one of the states adjacent to i, each with probability $1/K$, and remain at the state i itself with probability $1 - c(i)/K$. Then the transition probability p_{ij} is $1/K$ if $i \neq j$ and i, j are adjacent, 0 if $i \neq j$ and i, j are not adjacent, and $1 - c(i)/K$ if $i = j$. Clearly, now $\mathbf{P}$ is (symmetric and) doubly stochastic, and so the stationary distribution gives probability $1/N$ to each state provided the Markov chain is aperiodic. If $c(i) < K$ and so $p_{ii} > 0$ for at least one state i, that state and thus (noting that the Markov chain is irreducible) the entire Markov chain are aperiodic, and the distribution of the state after q steps tends toward the discrete uniform distribution as $q \to \infty$, whatever the initial state.

The difficulty in using the modification mentioned in the preceding paragraph is that one cannot get a good upper bound for the $c(i)$s. If K is too large compared to the $c(i)$s, then the p_{ii}s become close to 1, and the convergence to the uniform distribution will be too slow. The best K would, of course, be the maximum of $c(i)$s over all the states, increased by a small number to take care of periodicity. Since this exact maximum cannot be determined, we do the following: We estimate the maximum $c(i)$ using a pilot study and use the estimate increased by a small amount as the initial value of K. At any stage, if the current state is i, then we go from i to one of the states adjacent to i, each with probability $1/K$, and remain at i with probability $1 - c(i)/K$. If we move to state i', we update K by replacing it with $\max(K, c(i'))$. It can be shown that, in the limit, all states are equally probable.

One decision to make while using the Markov chain simulation method is the following: How long should the Markov chain be run to get a nearly random matrix? We do not have a definite answer, but it is known that any network can be obtained from any other network in t or fewer steps, where $t = \min(m, n(n-1) - m)$. Perhaps running the Markov chain for $2t$ to $3t$ steps will be enough to achieve a reasonably good level of mixing. Note that the network obtained in the $3t$th step is a nearly random network. To get another, we have to run the Markov chain again for $3t$ steps starting from the initial network.

3.6 OTHER MODELS

Model V

Recall that Model III.2 implicitly makes two assumptions: (i) For any fixed i, the probability that v_j is chosen by v_i is the same for all j, and (ii) the choices of different v_is are statistically independent. Assumption (i) was removed in Model III.3. In the present model, which is essentially that proposed by Katz and Powell (1956), we remove assumption (ii). This model is an incomplete model and stipulates that there is a common correlation τ between X_{ij} and X_{ji} for all i and j with $i \neq j$. $P(X_{ij} = 1)$ is still assumed to be $d_i/(n-1)$ for all $j \neq i$.

Now, from the definition of τ, we have

$$P(X_{ij} = X_{ji} = 1) = \frac{1}{(n-1)^2}\left(d_i d_j + \tau\sqrt{d_i d_j (n-1-d_i)(n-1-d_j)}\right)$$

and $E(s(G)|\tau)$ is the sum of the above probability over all i and j such that $i < j$. Equating $s(G)$ to its expected value, we get an estimate $\hat{\tau}$ of τ that can be taken as a measure of reciprocity. A positive value of $\hat{\tau}$ will be interpreted as indicating a tendency toward reciprocation and a negative value toward anti-reciprocation (j not going to i when i goes to j), while a value of 0 indicates neutrality with respect to reciprocation.

Note that when $d_i = d$ for all i, the above expression reduces to

$$P(X_{ij} = X_{ji} = 1) = \frac{d}{n-1}\left(\frac{d}{n-1} + \tau\frac{n-1-d}{n-1}\right).$$

Thus,

$$P(X_{ji} = 1 | X_{ij} = 1) = P(X_{ji} = 1) + \tau P(X_{ji} \neq 1).$$

Summing the above joint probability over i and j such that $i < j$, we get

$$E(s|\tau) = \frac{nd^2}{2(n-1)}(1-\tau) + \frac{nd}{2}\tau,$$

so

$$\hat{\tau} = \frac{2(n-1)s - nd^2}{nd(n-1-d)}.$$

This expression can be used as an approximation when d_is are nearly equal to d. Generalizing the above expression for $E(s|\tau)$ to

$$\frac{\sum\sum_{i<j} d_i d_j}{(n-1)^2}(1-\tau) + \frac{\sum_i d_i}{2}\tau,$$

(τ will not be the correlation coefficient and equation (6) will not hold now), Katz and Powell (1956) obtained the expression

$$\hat{\tau} = \frac{2(n-1)^2 s - S_1^2 + S_2}{(n-1)^2 S_1 - S_1^2 + S_2} \tag{3.40}$$

for $\hat{\tau}$ in the general case, where S_1 and S_2 are as defined in Model III.2. Notice that this reduces to the preceding expression when $d_i = d$ for all i.

For given d_is, the set of values $\hat{\tau}$ takes is contained in $[-1, 1]$. But $\hat{\tau}$ attains the value 1 only when s attains the value $\sum d_i/2$, and $\hat{\tau}$ attains the value -1 only when $d_i = (n-1)/2$ for all i and $s = 0$. When the d_is are all equal, τ is like an intraclass correlation coefficient.

The present model is incomplete since it does not even specify $P(X_{ij} = X_{ji} = X_{k\ell} = X_{\ell k} = 1)$. Thus, even the variance of s under the model cannot be computed.

Model VI

Wasserman and Faust (1999) discuss some useful models for understanding single relational data involving a given number of individuals in a social network of choice relation. We propose providing a description of these models, which are also termed *dyadic interaction models*. Some generalizations of these dyadic models have been recently studied in Em-ot, Tiensuwan, and Sinha (2008), and applications have also been made to real network data. We will provide a detailed description of the model and related data analysis results in Chapter 4 and also in Chapter 6 on graph-theoretic case studies.

As usual, we start with n individuals forming a network based on some form of choice relation between any pair of individuals. The possibilities of choice involving the pair of individuals i and j are

$$(0,0),\ (1,0),\ (0,1)\ \text{and}\ (1,1),$$

where the first component is 1 or 0 according to whether i chooses j or not, and the second component is 1 or 0 according to whether j chooses i or not.

The body of data arising from a network in the form of the adjacency matrix (or sociomatrix) of order $n \times n$ is referred to as X data. Note that $X_{ij} = 1$ or 0 according to whether there is an arc from i to j or not.

Y Array

In general, X data are not symmetric. For modeling purposes, we convert X data into what is called the "Y array." It is a symmetric matrix of order $2n \times 2n$ made up of a 2×2 submatrix $Y_{(i,j)}$ for each pair of vertices i and j.

The rows as well as the columns of $Y_{(i,j)}$ are designated 0 and 1 (instead of the usual 1 and 2). If $X_{ij} = r$ and $X_{ji} = s$, we put 1 in the (r, s)-cell of $Y_{(i,j)}$ and 0s in the other three cells. Clearly, $Y_{(j,i)}$ is the transpose of $Y_{(i,j)}$. The matrices $Y_{(i,j)}$ corresponding to the four possibilities for the pair (i, j) are shown below.

$$X_{ij} = 0, X_{ji} = 0: \qquad Y_{(i,j)} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$X_{ij} = 0, X_{ji} = 1: \qquad Y_{(i,j)} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$X_{ij} = 1, X_{ji} = 0: \qquad Y_{(i,j)} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$X_{ij} = 1, X_{ji} = 1: \qquad Y_{(i,j)} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

The contemplated model for the Y array dwells on specific expressions for the multinomial cell probabilities corresponding to the four cells designated by $(r, s), r = 0, 1;\ j = 0, 1$, for every pair of individuals (i, j). Let $P[(r, s); (i, j)]$ denote the probability that the cell (r, s) receives the value 1 in $Y_{(i,j)}$ (i.e., that $X_{ij} = r$ and $X_{ji} = s$). Then the model stipulates the following:

$$\begin{aligned}
\log P[(0,0);(i,j)] &= \lambda_{ij},\\
\log P[(1,0);(i,j)] &= \lambda_{ij}+\theta+\alpha_i+\beta_j,\\
\log P[(0,1);(i,j)] &= \lambda_{ij}+\theta+\alpha_j+\beta_i,\\
\log P[(1,1);(i,j)] &= \lambda_{ij}+2\theta+\alpha_i+\beta_j+\alpha_j+\beta_i+(\alpha\beta),
\end{aligned} \tag{3.41}$$

where the logarithm is the natural logarithm (w.r.t. the base e).

In the above, for every pair (i, j), λ_{ij} is a normalizing constant since the four probability expressions have to add up to unity. This yields

$$\lambda_{ij} = -\log\left(1+e^{\theta+\alpha_i+\beta_j}+e^{\theta+\alpha_j+\beta_i}+e^{2\theta+(\alpha_i+\beta_j)+(\alpha_j+\beta_i)+(\alpha\beta)}\right). \tag{3.42}$$

The other parameters are interpreted as follows:

1. θ represents what may be termed the overall "choice" parameter (like the grand mean in analysis of variance [ANOVA]), and for each choice (by either i or j or by both), its presence in the model expression for probability is thus incorporated.

2. α represents the "expansiveness" or "out-degree" aspect and β the "popularity" or in-degree aspect of an individual, and these are individual specific in general terms.

3. $(\alpha\beta)$ represents the "reciprocity" aspect of the pair of individuals concerned, and it is regarded as a global phenomenon (so that its dependence on specific individuals' features is ruled out).

More general models are discussed in Wasserman and Faust (1999).

Note that while the in-degree and/or out-degree aspects are assumed to be individual specific, the reciprocity aspect is taken to be a "global" phenomenon and not pairwise individual specific. The same is true of θ as a parameter indicating the "global" aspect of choices of individuals in the network. It receives weight 2 for the last case when both the individuals choose each other. We must also note that there are altogether $n(n-1)$ ordered pairs of individuals indexed like (i, j) in terms of X data, but in terms of Y array, we have to consider only $n(n-1)/2$ *dyads*—that is, unordered pairs or, equivalently, ordered pairs (i, j) with $i < j$—since $Y_{(i,j)}$ determines $Y_{(j,i)}$. Furthermore, in the model described above, there are altogether $1+n+n+1=2(n+1)$ parameters. As a matter of convention, it is assumed that

$$\sum \alpha_i = \sum \beta_i = 0$$

so that, in effect, there are $2n$ parameters to be estimated. For a given pair of individuals, the likelihood function is readily expressible in terms of the multinomial (4-cell) probabilities, using the indicator functions. Hence, the joint log-likelihood function for all such *unordered pairs* of individuals can be written down in a routine manner. The main problem in finding maximum likelihood estimates of the parameters lies in the dependence of the λ_{ij}s on the others—namely, $\theta, \alpha_i, \alpha_j, \beta_i, \beta_j$, and $(\alpha\beta)$. Therefore, the likelihood equations are not analytically tractable, and one has to take recourse to statistical computing.

W Array

Having described a general model as above, we can now focus the discussion on some simplified models that are motivated by considerations of possible "grouping" among the individuals. In a sociological context, these groupings may be prompted by family size, occupation, caste, kinship, and the like. If we consider one such "external factor," then the individuals may be classified into several disjoint categories, and the category-specific out-degree and in-degree parameters may be more appropriate to use. That means we can dispense with individual-specific α_i and β_j parameters and replace them by those associated with their groups. Thus, for two groups or categories I and II (like small/large family size), we may use α_I and α_{II} along with the constraint $n_I\alpha_I + n_{II}\alpha_{II} = 0$, where n_I and n_{II} are the group sizes. The same is true of the βs, the in-degree parameters. Of course, m and $(\alpha\beta)$ remain unaltered. This leads to what has been referred to as a W array. It may be noted that under the W array, the number of parameters reduces to $1 + (C-1) + (C-1) + 1 = 2C$, where C is the number of categories induced by the external factor.

For the W array with two categories, the model description is given by

$$
\begin{aligned}
\log P[(0,0);(i,j)] &= \lambda_{I,I},\\
\log P[(1,0);(i,j)] &= \lambda_{I,I} + \theta + \alpha_I + \beta_I,\\
\log P[(0,1);(i,j)] &= \lambda_{I,I} + \theta + \alpha_I + \beta_I,\\
\log P[(1,1);(i,j)] &= \lambda_{I,I} + 2\theta + 2\alpha_I + 2\beta_I + (\alpha\beta)
\end{aligned}
\tag{3.43}
$$

when the two individuals are *both* in Category I. Similar descriptions apply for the cases: Both are in Category II or one is in Category I, while the other is in Category II. For example, if i and j belong to different categories, we may assume without loss of generality that i belongs to Category I and j to Category II (since we have to consider only unordered pairs of individuals), and then

$$\begin{aligned}
&\log P[(0,0);(i,j)] = \lambda_{I,II},\\
&\log P[(1,0);(i,j)] = \lambda_{I,II} + \theta + \alpha_I + \beta_{II},\\
&\log P[(0,1);(i,j)] = \lambda_{I,II} + \theta + \alpha_{II} + \beta_I,\\
&\log P[(1,1);(i,j)] = \lambda_{I,II} + 2\theta + \alpha_I + \beta_{II} + \alpha_{II} + \beta_I + (\alpha\beta).
\end{aligned} \tag{3.44}$$

Moreover, for this model involving two groups, there are altogether six parameters—namely, θ, α_I, α_{II}, β_I, β_{II}, and $(\alpha\beta)$.

V Array

There is yet another simplified version of the *Y* array. For example, in situations where individual αs and βs are insignificant or sufficiently small compared to the other parameters, we can ignore them in the model description. This is referred to as a *V* array, and the model with only two parameters reduces to

$$\begin{aligned}
&\log P[(0,0);(i,j)] = \lambda,\\
&\log P[(1,0);(i,j)] = \lambda + \theta,\\
&\log P[(0,1);(i,j)] = \lambda + \theta,\\
&\log P[(1,1);(i,j)] = \lambda + 2\theta + (\alpha\beta).
\end{aligned} \tag{3.45}$$

Explicit expression for the maximum likelihood estimate of $(\alpha\beta)$ can be obtained under the present model (see Rao & Bandyopadhyay, 1987). Rao and Bandyopadhyay (1987) also discuss the special case of this model obtained by dropping the parameter θ and keeping only the parameter $(\alpha\beta)$.

Each of the parameters θ, α_is, β_is, and $(\alpha\beta)$ can be estimated, for example, by the method of maximum likelihood. However, there are no closed formulae for the estimates, and they have to be computed by iterative procedures (for details, see Wasserman & Faust, 1999). We will discuss some aspects of model fitting and model validation in Chapter 4, wherein the problem of estimating these parameters will also be briefly addressed. Illustrative examples will be discussed in Chapter 6 on case studies.

Model VII

Model VII, due to Holland and Leinhardt (1981), assumes that the probability of obtaining any particular network is

$$\text{const.}\exp\left\{\theta m + \sum_{i=1}^{n}\alpha_i d_i + \sum_{i=1}^{n}\beta_i e_i + \rho s\right\},$$

where θ, ρ, α_is, and β_is are real valued parameters such that $\sum \alpha_i = 0$ and $\sum \beta_i = 0$. Note that only n is fixed, and m, d_is, and e_is are variables.

Under the present model, the distributions of (X_{ij}, X_{ji}), and $(X_{k\ell}, X_{\ell k})$ are independent if i, j, k, and ℓ are distinct. Thus, it allows for correlation between X_{ij} and X_{ji}, but disjoint pairs are independent. Here, θ represents density (or overall choice), α_is represent the differential expansiveness of the vertices, β_is represent the differential popularity of the vertices, and ρ represents overall reciprocity. It can be verified that

$$\frac{P(X_{ji}=1|X_{ij}=1)}{P(X_{ji}=0|X_{ij}=1)} \Big/ \frac{P(X_{ji}=1|X_{ij}=0)}{P(X_{ji}=0|X_{ij}=0)} = e^{\rho}$$

is independent of i and j. Thus, ρ is a log-odds ratio. One can get special cases of the present model corresponding to Models II.2 and III.2 as follows: For II.2, drop the terms containing d_is, e_is, and s. For III.2, drop the terms containing e_is and s. On the other hand, it turns out that the dyadic model (described in the preceding section and expanded by Wasserman & Faust, 1999) encompasses this model as the "joint likelihood of the parameters" based on the entire body of dyadic interaction data on all pairs of vertices with the identification of ρ as $\alpha\beta$.

As mentioned before, each of the parameters θ, ρ, α_is, and β_is can be estimated, for example, by the method of maximum likelihood. However, there are no closed formulae for the estimates, and they have to be computed by iterative procedures. For details, see Holland and Leinhardt (1981) and Wasserman and Faust (1999).

REFERENCES

Achuthan, N., Rao, S. B., & Rao, A. R. (1984). The number of symmetric edges in a digraph with prescribed out-degrees. In *Combinatorics and applications: Proceedings of the Seminar in Honour of Professor S. S. Shrikhande* (pp. 8–20). Calcutta: Indian Statistical Institute.

Berge, C. (1973). *Graphs and hypergraphs*. Amsterdam: North-Holland.

Em-ot, P., Tiensuwan, M., & Sinha, B. K. (2008). Some aspects of stochastic modeling of dyadic relations in social networks: Theory and applications. *Journal of Statistical Theory and Applications, 7,* 303–322.

Feller, W. (1968). *An Introduction to probability theory and its applications* (Vol. 1, 3rd ed.). New York: John Wiley.

Fulkerson, D. R. (1966). A network flow computation for project cot curves. *Management Science, 7,* 167–178.

Harary, F. (1988). *Graph theory*. New Delhi, India: Narosa.

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

Katz, L., & Powell, J. H. (1954). The number of locally restricted directed graphs. *Proceedings of the American Mathematics Society, 5,* 621–626.

Katz, L., & Powell, J. H. (1956). Measurement of tendency towards reciprocation of choice. In J. L. Moreno (Ed.), *Sociometry and the science of man* (pp. 403–409). Beacon, NY: Beacon House.

Katz, L., & Wilson, T. R. (1956). The variance of the number of mutual choices in sociometry. *Psychometrika, 21,* 299–304.

Pramanik, P. (1994, January). *Generating random* (0, 1)*-matrices with given marginals*. Paper presented at the Third International Conference on Lattice Path Combinatorics, New Delhi, India.

Rao, A. R. (1984). *Bounds for the number of symmetric edges in a digraph with specified out-degrees and in-degrees*. Unpublished manuscript, Indian Statistical Institute.

Rao, A. R. (2002, December). *The number of reachable pairs in a digraph*. Paper presented at the R. C. Bose Centenary Symposium on Discrete Mathematics and Applications, Indian Statistical Institute, Kolkata, India.

Rao, A. R., & Bandyopadhyay, S. (1987). Measures of reciprocity in a social network. *Sankhyā*, Series A, 49, 141–188.

Rao, A. R., Jana, R., & Bandyopadhyay, S. (1996). A Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā, Series A, 58,* 225–242.

Ryser, H. J. (1963). *Combinatorial mathematics*. New York: John Wiley.

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

Sukhatme, P. V. (1938). On bipartitional functions. *Philosophical Transactions of the Royal Society of London, 237,* 375–409.

Wasserman, S., & Faust, K. (1999). *Social network analysis: Methods and applications.* Cambridge, UK: Cambridge University Press.