# Random graph models of social networks

M. E. J. Newman
*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

D. J. Watts
*Department of Sociology, Columbia University, 1180 Amsterdam Avenue, New York, NY 10027*

S. H. Strogatz
*Department of Theoretical and Applied Mechanics, Cornell University, Ithaca, NY 14853–1503*
(Dated: June 29, 2001)

We describe some new exactly solvable models of the structure of social networks, based on random graphs with arbitrary degree distributions. We give models both for simple unipartite networks, such as acquaintance networks, and bipartite networks, such as affiliation networks. We compare the predictions of our models to data for a number of real-world social networks and find that in some cases the models are in remarkable agreement with the data, while in others the agreement is poorer, perhaps indicating the presence of additional social structure in the network that is not captured by the random graph.

## I. INTRODUCTION

A social network is a set of people or groups of people, "actors" in the jargon of the field, with some pattern of interactions or "ties" between them [1, 2]. Friendships among a group of individuals, business relationships between companies, and intermarriages between families are all examples of networks which have been studied in the past. Network analysis has a long history in sociology, the literature on the topic stretching back at least half a century to the pioneering work of Rapaport, Harary, and others in the 1940s and 1950s. Typically network studies in sociology have been data-oriented, involving empirical investigation of real-world networks followed, usually, by graph theoretical analysis often aimed at determining the centrality or influence of the various actors.

More recently, following a surge in interest in network structure among mathematicians and physicists, partly as a result of research on the Internet and the World-Wide Web, another body of research has investigated the statistical properties of networks and methods for modeling networks either analytically or numerically [3, 4]. One important and fundamental result which has emerged from these studies concerns the numbers of ties that actors have to other actors, their so-called "degree." It has been found that in many networks the distribution of actors' degrees is highly skewed, with a small number of actors having an unusually large number of ties. Simulations and analytic work have suggested that this could have an impact on the way in which communities operate, including the way information travels through the network and the robustness of networks to removal of actors [5–7]. In this paper we describe some new models of social networks which allow us to explore directly the effects of varying degree distributions.

## II. EMPIRICAL DATA

Before discussing our models, we first describe briefly some of the empirical results about real-world social networks which has motivated our work.

Recent work on social networks within mathematics and physics has focussed on three distinctive features of network structure. The first of these is the "small world" effect, which was highlighted in early work by Pool and Kochen [8] and by Stanley Milgram [9]. In his now-classic 1967 paper [9], Milgram described an experiment he performed involving letters that were passed from acquaintance to acquaintance, from which he deduced that many pairs of apparently distant people are actually connected by a very short chain of intermediate acquaintances. He found this chain to be of typical length only about six, a result which has passed into folklore via John Guare's 1990 play *Six Degrees of Separation* [10]. It has since been shown that many networks have a similar small-world property [11–14].

It is worth noting that the phrase "small world" has been used to mean a number of different things. Early on, sociologists used the phrase both in the conversational sense of two strangers who discover that they have a mutual friend—i.e., that they are separated by a path of length two—and to refer to any short path between individuals [8, 9]. Milgram talked about the "small world problem," meaning the question of how two people can have a short connecting path of acquaintances in a network that has other social structure such as insular communities or geographical and cultural barriers. In more recent work, Watts and Strogatz [11] have used the phrase "small world network" to mean a network which exhibits this combination of short paths and social structure, the latter being defined in their case in terms of network clustering (see below). The reader may find it helpful to bear these different definitions in mind when reading this and other papers on this topic.

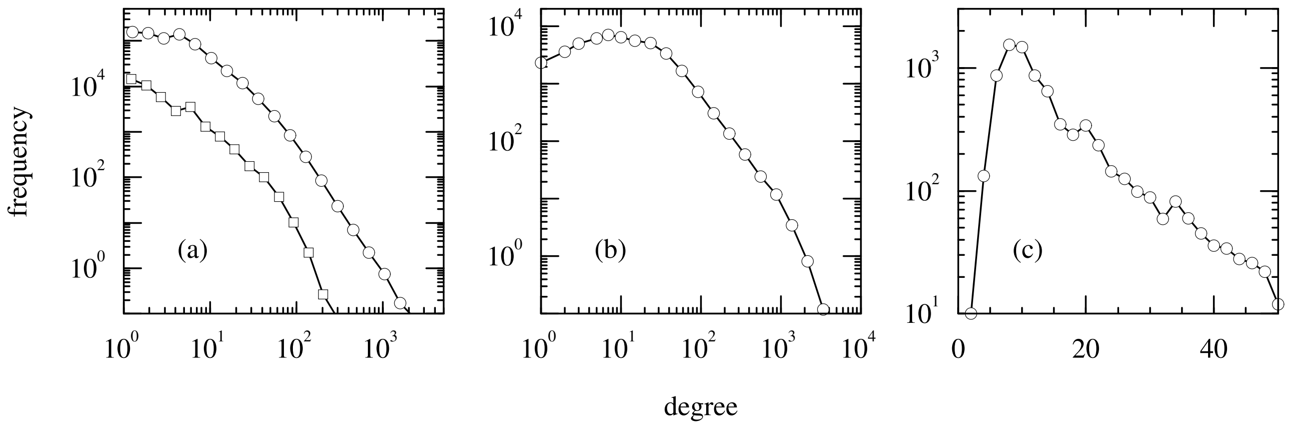


FIG. 1: Degree distributions for three different types of networks: (a) scientific collaboration networks of biologists (circles) and physicists (squares); (b) a collaboration network of movie actors; (c) network of directors of Fortune 1000 companies. Note that panel (c) has a linear horizontal axis, while all other axes are logarithmic. Solid lines between points are merely a guide to the eye. Panels (a) and (b) after Newman [14] and Amaral *et al.* [13], respectively. Data for panel (c) kindly provided by G. Davis.

The second property of social networks which has been emphasized in recent work is clustering. In a paper in 1998, Watts and Strogatz [11] showed that in many real-world networks the probability of a tie between two actors is much greater if the two actors in question have another mutual acquaintance, or several. To put that another way, the probability that two of your friends know one another is much greater than the probability that two people chosen randomly from the population know one another. Watts and Strogatz defined a "clustering coefficient," usually denoted $C$, which is the probability that two acquaintances of a randomly chosen person are themselves acquainted. They showed for a variety of networks that this clustering coefficient took values anywhere from a few percent to forty or fifty percent, and other studies have since shown similar results for other networks [14, 15]. In many cases this makes the probability of acquaintance between people several orders of magnitude greater if they have a common friend than if they do not.

The third of our three properties of networks is perhaps the most important for the work described in this paper, and was mentioned in the introduction. It is the property of having a skewed degree distribution, which has been particularly emphasized in the work of Albert, Barabási, and co-workers [12, 16]. In Fig. 1 we show histograms of degree distributions, i.e., number of actors having a given degree, for three different types of networks, all of them, arguably, social networks. The networks shown are as follows.

a) Scientific collaboration networks [14, 17]: networks in which the actors are scientists in various fields and the ties between them are collaborations, defined as coauthorship of one or more scientific papers during the period of the study (degree = number of collaborators of a scientist).

b) Movie actor collaborations [11, 13]: a network in which the actors are, well, actors—movie actors in this case—and a tie between two of them represents appearance in the same movie (degree = number of other actors with whom an actor has co-starred).

c) Company directors [18, 19]: a network in which the actors are company directors of companies in the 1999 Fortune 1000 (the one thousand US companies with the largest revenues in 1999). A tie between two directors indicates that they sat on the same board together (degree = number of others with whom a director sits on boards).

In the first two of these networks, the degree distribution has a highly skewed form, approximately obeying a power law for a part of its range (a straight line on the logarithmic scales used), although having an apparently exponential cutoff for very high values of the degree [13]. In the third network, the distribution is much less skewed, having a sharp peak around degree 10, and a fast (approximately exponential) decay in the tail. One possible explanation for the difference between the first two cases and the third is that maintenance of ties in the third network, the network of company directors, has a substantial cost associated with it. It takes continual work to be a company director. Collaboration between scientists or movie actors on the other hand carries only a one-time cost, the time and effort put into writing a paper or making a movie, but the tie gained is (by the definition used here) present indefinitely thereafter. This difference may put a sharper limit on the number of directorships a person can hold than on numbers of collaborators.

Recent research on networks has focused a lot of attention on those networks with skewed degree distribu-

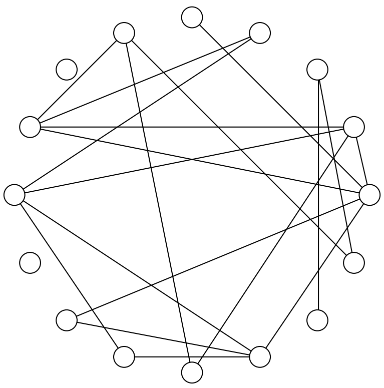

FIG. 2: An example of a standard random graph of the type first discussed by Erdős and Rényi [23]. In this case the number of vertices $N$ is 16 and the probability $p$ of an edge is $\frac{1}{7}$.

tions [3, 4, 12, 13, 20–22], and we will consider these in the present paper also. However, the methods and models we will describe are not restricted to this case. As we will show our models can be used to mimic networks with any desired degree distribution.

## III. RANDOM GRAPHS WITH ARBITRARY DEGREE DISTRIBUTIONS

In 1959, Paul Erdős and Alfréd Rényi published a seminal paper in which they introduced the concept of a random graph [23]. A random graph is simple to define. One takes some number $N$ of nodes or "vertices" and places connections or "edges" between them, such that each pair of vertices $i, j$ has a connecting edge with independent probability $p$. We show an example of such a random graph in Fig. 2. This is one of the simplest models of a network there is, and is certainly the best studied; the random graph has become a cornerstone of the discipline known as discrete mathematics and many hundreds of papers have discussed its properties. However, as a model of a real-world network it has some serious shortcomings. Perhaps the most serious is its degree distribution, which is quite unlike those seen in most real world networks.

Consider a vertex in a random graph. It is connected with equal probability $p$ with each of the $N-1$ other vertices in the graph, and hence the probability $p_k$ that it has degree exactly $k$ is given by the binomial distribution:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \qquad (1)$$

Noting that the average degree of a vertex in the network is $z = (N-1)p$ we can also write this as

$$p_k = \binom{N-1}{k} \left[\frac{z}{N-1-z}\right]^k \left[1 - \frac{z}{N-1}\right]^{N-1} \simeq \frac{z^k}{k!}\mathrm{e}^{-z}, \qquad (2)$$

where the last approximate equality becomes exact in the limit of large $N$. We recognize this distribution as the Poisson distribution. A large random graph has a Poisson degree distribution. This makes the random graph a poor approximation to the real-world networks discussed in the previous section, with their highly skewed degree distributions. On the other hand, the random graph has many desirable properties, particularly the fact that many features of its behavior can be calculated exactly. This leads us to ask an obvious question: is it possible to create a model which matches real-world networks better but is still exactly solvable? We show now that it is.

Suppose that we want to make a model of a large network for which we know the degree distribution but nothing else. That is, we are given the (properly normalized) probabilities $p_k$ that a randomly chosen vertex in the network has degree $k$. We can make a model network with this same degree distribution using the following algorithm, which is due to Molloy and Reed [24]. We take a number $N$ of vertices, and we assign to each a number $k$ of "stubs" or ends of edges, where $k$ is a random number drawn independently from the distribution $p_k$ for each vertex. Now we choose those stubs randomly in pairs and join them up to form edges between the vertices. This procedure will produce a graph with exactly the desired degree distribution, but which is in all other respects random. To put it another way, we have generated a graph which is drawn uniformly at random from the set of graphs with the given degree distribution. Given that the degree distribution was the only information that we had about the network in question, this is the appropriate thing to do.

(The algorithm above has one small problem: if the number of stubs generated is odd, we cannot match them all up in pairs. This is easily corrected however: if the number is found to be odd, we throw one vertex away and generate a new one from the distribution $p_k$, repeating until the number of stubs is even.)

This then is our simplest model for a social network.

## IV. EXACT RESULTS

It turns out that many properties of the network model described above are exactly solvable in the limit of large network size. The crucial trick for finding the solution is that instead of working directly with the degree distribution $p_k$, we work with a "generating function" $G_0(x)$, which is defined as

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \qquad (3)$$

This function encapsulates all of the information in $p_k$, but does so in a form which turns out to be easier to work with than $p_k$ itself. Notice for example that the average degree $z$ of a vertex in the network is given simply in terms of a derivative of $G_0$:

$$z = \sum_k k p_k = G_0'(1). \tag{4}$$

Notice also that the normalization condition on $p_k$ has a simple expression in terms of the generating function: if $p_k$ is properly normalized then $G_0(1) = 1$.

Here we will not go into all the details of our derivations, but give a summary of the important results. The reader in search of mathematical nitty-gritty should consult Ref. 15.

The most striking property of our model networks is that they exist in two different regimes. Depending on the exact distribution $p_k$ of the degrees of vertices they may either be made up of many small clusters of vertices connected together by edges, also called "components," or they may contain a "giant component"—a group of connected vertices which fills a significant portion of the whole network and whose size scales up with the size of the whole network—in addition to a number of small components. The fraction $S$ of the network which is filled by the giant component is given by

$$S = 1 - G_0(u), \tag{5}$$

where $u$ is the smallest non-negative real solution of

$$zu = G_0'(u), \tag{6}$$

with $z$ given by Eq. (4). (This result is not new to our work: an equivalent formula has been derived previously by different methods; see Ref. 25.) For some distributions $p_k$, Eqs. (5) and (6) give $S = 0$, which indicates that there is no giant component.

The average size of components in the network, excluding the giant component if there is one, is

$$\langle s \rangle = 1 + \frac{z^2 u^2}{G_0(u)[z - G_0''(u)]}. \tag{7}$$

To give a feeling for what these results mean, consider the following degree distribution:

$$p_k = \begin{cases} 0 & \text{for } k = 0 \\ Ck^{-\tau} \mathrm{e}^{-k/\kappa} & \text{for } k \geq 1. \end{cases} \tag{8}$$

This is a distribution of the form seen in the first two panels of Fig. 1: a power-law distribution charaterized by the exponent $\tau$, with an exponential cutoff characterized by the cutoff length $\kappa$. The constant $C$ is fixed by the requirement that the distribution be normalized $\sum_k p_k = 1$, which gives $C = [\mathrm{Li}_n(\mathrm{e}^{-1/\kappa})]^{-1}$, where $\mathrm{Li}_n(x)$ is the $n^{\text{th}}$ polylogarithm of $x$. Thus

$$p_k = \frac{k^{-\tau} \mathrm{e}^{-k/\kappa}}{\mathrm{Li}_\tau(\mathrm{e}^{-1/\kappa})} \qquad \text{for } k \geq 1. \tag{9}$$

Substituting into Eq. (3), we then get

$$G_0(x) = \frac{\mathrm{Li}_\tau(x\mathrm{e}^{-1/\kappa})}{\mathrm{Li}_\tau(\mathrm{e}^{-1/\kappa})}. \tag{10}$$

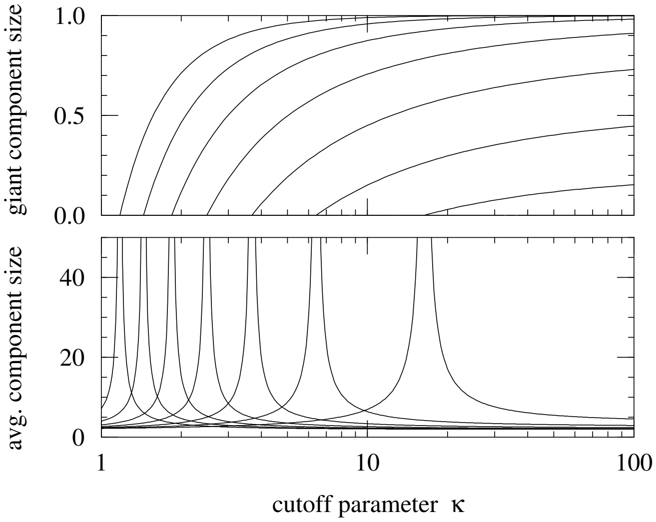


FIG. 3: Size $S$ of the giant component (top) and average size $\langle s \rangle$ of clusters excluding the giant component (bottom) for graphs with the degree distribution given in Eq. (9), as a function of the cutoff parameter $\kappa$. The curves are for, left to right, $\tau = 0.6$ to $3.2$ in steps of 0.4.

We can now use this function in Eqs. (5–7) to find the size of the giant component and the average component size for graphs of this type. The results are shown in Fig. 3.

The figure shows $S$ and $\langle s \rangle$ as a function of the cutoff parameter $\kappa$ for a variety of different values of the exponent $\tau$. The transition at which the giant component appears is clearly visible in the top panel for each curve, and occurs at a value of $\kappa$ which gets larger as $\tau$ gets larger. The average cluster size $\langle s \rangle$ diverges at this transition point, as seen in the bottom panel of the figure.

The existence, or not, of a giant component in the network has important implications for social networks. If for example information spreads on a network by person-to-person communication, it can only get from person A to person B if there is at least one connected path of individuals from A to B through the network. The components in a network are precisely those sets of individuals who have such a path between them, and hence can communicate with one another in this way. If there is no giant component in a network, then all components are small and communication can only take place within small groups of people of typical size $\langle s \rangle$. If on the other hand there is a giant component, then a large fraction of the network can all communicate with one another, and the number $S$ is this fraction.

Looking at Eq. (7) we see that the divergence in $\langle s \rangle$ occurs when $G_0''(u) = z$. We also know that $S = 0$ at this point, and using Eq. (5) and the fact that $G_0(1) = 1$ always, we conclude that the transition point at which the giant component appears is given by

$$G_0''(1) = z. \tag{11}$$

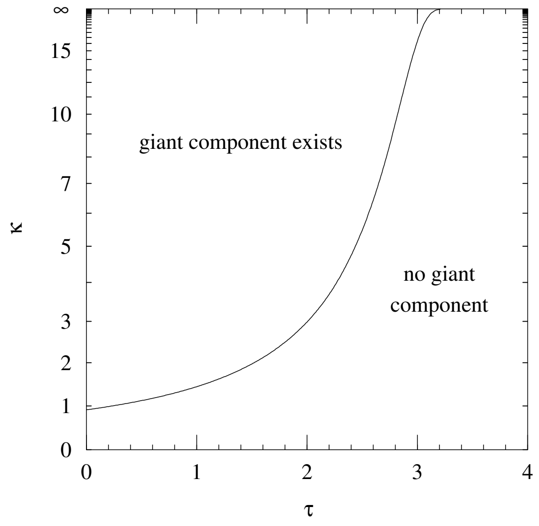


FIG. 4: Phase diagram for networks with the skewed degree distribution defined in Eq. (9). The solid line marks the boundary between the region in which a giant component exists and the one in which it does not.

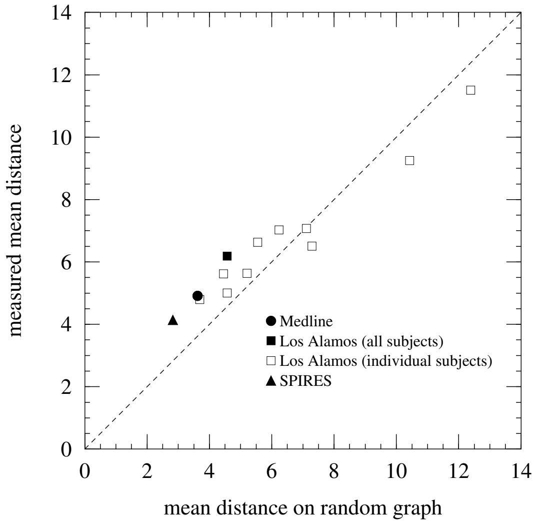


FIG. 5: Mean distance between vertices in 13 scientific collaboration networks, from the theoretical prediction, Eq. (14), and from direct measurements. If theory and measurement agreed perfectly, all points would lie on the dotted line. After Newman [17].

As an example of this we show in Fig. 4 the resulting "phase diagram" for the class of networks defined by Eq. (9). This plot shows which regions of the $\tau$–$\kappa$ plane contain a giant component and which do not. Two special points worthy of note in this figure are the points at which the solid line marking the phase boundary intersects the axes. At one end it intersects the line $\tau = 0$ at the point $\kappa = [\log 3]^{-1} = 0.9102\ldots$, implying that when $\kappa$ is below this value a giant component can never exist, regardless of the value of $\tau$. At the other end it intersects the line $\kappa = \infty$ at a value of $\tau$ which is the solution of $\zeta(\tau - 2) = 2\zeta(\tau - 1)$, or around $\tau = 3.4788\ldots$, implying that for values of $\tau$ larger than this, a giant component can never exist, regardless of the value of $\kappa$. The second of these results was derived previously by Aiello *et al.* using a different method [26].

Almost all networks found in society and nature appear to be well inside the region in which the giant component exists; networks with no obvious giant component are rare. This may be a tautology however: it is possible that it rarely occurs to researchers to consider a network representation of a system which is not heavily interconnected.

We can also show that our networks have short average path lengths between vertices, path lengths that increase logarithmically with the size $N$ of the network. We find that the average number $z_m$ of vertices a distance $m$ steps away from a given vertex is given recursively by

$$z_m = \frac{G_0''(1)}{G_0'(1)} z_{m-1}, \tag{12}$$

and hence that

$$z_m = \left[\frac{z_2}{z_1}\right]^{m-1} z_1, \tag{13}$$

where $z_1 = z$ is synonymous with the average degree of a vertex and $z_2$ is the average number of second neighbors of a vertex. Thus, if we know these two numbers for a network, then we can predict the average number of neighbors any distance away from a given vertex.

To calculate typical path lengths in the network, we now observe that when the number of vertices $z_\ell$ a distance $\ell$ away from a given vertex is equal to the total number of vertices in the whole network, then $\ell$ is roughly equal to the typical distance between all pairs of vertices. Substituting $m \to \ell$ and $z_\ell \to N$ in Eq. (13) and rearranging, we then get

$$\ell = \frac{\log(N/z_1)}{\log(z_2/z_1)} + 1. \tag{14}$$

Thus the typical distance between vertices is indeed increasing only logarithmically with $N$.

As a demonstration of this result, consider Fig. 5, in which we show the mean distance between vertices in thirteen actual networks of collaborations between scientists, as described in Section II, plotted against the values of the same quantities predicted by Eq. (14). If theoretical and empirical values agreed perfectly, all the points in the figure would fall on the dotted diagonal line. As the figure shows, agreement is not perfect, but nonetheless sufficiently good to give us some confidence in the theory.

## V. AFFILIATION NETWORKS AND BIPARTITE GRAPHS

One of the biggest problems in studying social networks is the presence of uncontrolled biases in the empirical data. Studies of acquaintance networks and similar social networks are usually carried out either by interviewing participants or by circulating questionnaires, asking actors to identify others with whom they have ties of one sort or another. Studies of this kind have taught us much about the structure of society, but the experimental method has some problems. First, the data derived are limited in number, since it takes a lot of work to compile a data set of any substantial size, and practical studies have been limited mostly to a few tens or hundreds of actors. Second, there are inevitably large subjective biases in the data obtained, deriving from variations in the view of the respondents about what constitutes a tie and how strong those ties are.

There is however one type of social network that in many cases avoids both of these shortcomings, the so-called affiliation network. An affiliation network is a network in which actors are joined together by common membership of groups or clubs of some kind. Examples which have been studied in the past include networks of individuals joined together by common participation in social events [27] and CEOs of companies joined by common membership of social clubs [28]. The collaboration networks of scientists and movie actors and the network of boards of directors introduced in Section II are also affiliation networks, in which the groups to which actors belong are the groups of authors of a scientific paper, the groups of actors appearing in a single movie, or the groups of directors on a single board. Because membership of groups can frequently be established from membership lists or other resources, studies of these networks need not rely on interviews or questionnaires, and this makes possible the construction of much larger and more accurate networks than in for traditional social network studies. In the case of the networks of scientists, for example, scientists' coauthorship of papers may be recorded in bibliographic databases, and these databases can then be used to reconstruct the collaboration network [14].

Often affiliation networks are represented simply as unipartite graphs of actors joined by undirected edges—two company directors who sit on a common board, for example, being connected by an edge. However, this representation misses out much of the interesting structure of affiliation networks. Affiliation networks are, at heart, bipartite structures: the information they contain is most completely represented as a graph consisting of two kinds of vertices, one representing the actors and the other the groups. Edges then run only between vertices of unlike kinds, connecting actors to the groups to which they belong. The bipartite and unipartite representations of a small example network are illustrated in Fig. 6.

We can model affiliation networks using machinery very similar to that introduced in Section IV. For an

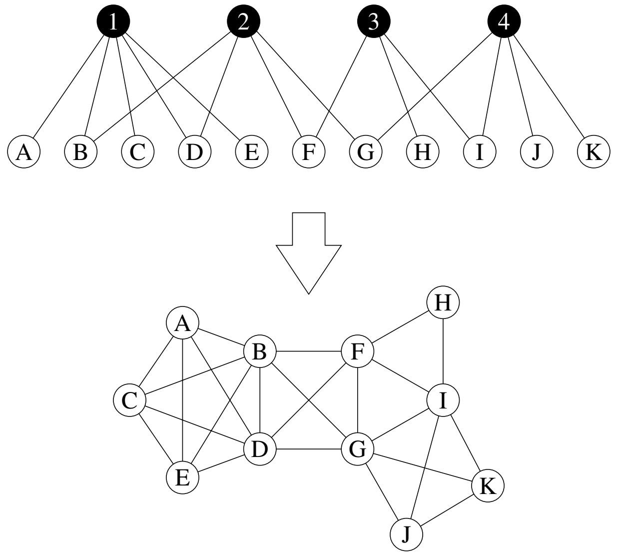


FIG. 6: (a) A bipartite network. One can imagine the vertices A to K as being, for example, company directors, and vertices 1 to 4 as being boards that they sit on, with lines joining each director to the appropriate boards. (b) The unipartite projection of the same network, in which two directors are connected by an edge if they sit on a board together.

affiliation network, there are two different degree distributions. To be concrete, we will describe the developments in terms of company directors and boards, but our results are applicable to any affiliation network. The two degree distributions then are the distribution of the number of boards that directors sit on, and the number of directors who sit on boards.

As a model for bipartite networks we consider a random bipartite graph in which the two types of vertices have the correct degree distributions, but again vertices (of unlike kinds) are paired up at random to create the model network. To treat such networks mathematically we define two generating functions, one for each of the two degree distributions. If we denote the probability that a director sits on $j$ boards by $p_j$ and the probability that a board has $k$ directors on it by $q_k$, then the two functions are

$$f_0(x) = \sum_j p_j x^j, \qquad g_0(x) = \sum_k q_k x^k. \tag{15}$$

From these we can then define a further function

$$G_0(x) = f_0\big(g_0'(x)/g_0'(1)\big), \tag{16}$$

which is the generating function for the number of neighbors of a director in the unipartite projection of the affiliation network pictured in Fig. 6b. This function plays exactly the same role as the function with the same name in Section IV, and essentially all of the same results apply. The average number of codirectors of a vertex in the network is $z = G_0'(1)$. The affiliation network shows

| network | clustering $C$ theory | clustering $C$ actual | average degree $z$ theory | average degree $z$ actual |
|---|---|---|---|---|
| company directors | 0.590 | 0.588 | 14.53 | 14.44 |
| movie actors | 0.084 | 0.199 | 125.6 | 113.4 |
| physics | 0.192 | 0.452 | 16.74 | 9.27 |
| biomedicine | 0.042 | 0.088 | 18.02 | 16.93 |

TABLE I: Summary of results of the analysis of four affiliation networks.

a phase transition at which a giant component appears at a point given by Eq. (11). The size of the giant component is given by Eqs. (5) and (6), the average size of other components is given by Eq. (7), and the typical vertex–vertex distance through the network is given by Eq. (14). (In fact, we implicitly made use of this last result in constructing Fig. 5, since the networks depicted in that figure are really affiliation networks.)

However, there are other results which are peculiar to bipartite networks. For example, the clustering coefficient $C$ which was discussed in Section II is asymptotically zero for the unipartite random graphs of Section IV—specifically, $C \sim N^{-1}$ for all degree distributions and hence goes to zero as $N \to \infty$. This is not true however for bipartite random graphs. To see this consider the following expression for the clustering coefficient (which is one of a number of ways it can be written):

$$C = \frac{3\times \text{ number of triangles on the graph}}{\text{number of connected triples of vertices}}. \quad (17)$$

Here "triangles" are trios of vertices each of which is connected to both of the others, and "connected triples" are trios in which at least one is connected to both the others. The factor of 3 in the numerator accounts for the fact that each triangle contributes to three connected triples of vertices, one for each of its three vertices. With this factor of 3, the value of $C$ lies strictly in the range from zero to one. Looking again at Fig. 6b, we see that there are many triangles in the network of directors, triangles which arise whenever there are three or more directors on a single board. Thus as long as there is a significant density of such boards in the network, the value of $C$ will be non-zero in the limit of large graph size. In fact, it turns out that the clustering coefficient can be expressed simply in terms of the generating functions $g_0$ and $G_0$ thus:

$$C = \frac{M}{N}\frac{g_0'''(1)}{G_0''(1)}, \quad (18)$$

where $M$ is the total number of boards of directors in the network and $N$ the total number of directors.

In Table I we compare the predictions of our method for $C$ and for average numbers of codirectors/collaborators $z$ against actual measurements for the four affiliation networks of Fig. 1: boards of directors for the 1999 Fortune 1000 [19]; collaborations of movie actors taken from the Internet Movie Database [29]; and two networks of scientific collaborations between 1995 and 1999, one in biology and medicine, and one in physics [14]. In the calculations, the degree distributions $p_j$ and $q_k$ used to define the generating functions were taken directly from the actual networks; that is we created networks that had degree distributions identical to those of the real-world networks they were supposed to mimic, but which were in all other respects entirely random.

As the table shows, our theory is remarkably precise for the network of boards of directors. Both $C$ and $z$ are predicted to within 1%. For the other networks the results are not as good. The average number of collaborators is predicted with moderate accuracy, but the values for the clustering coefficient, while they are of the right order of magnitude, appear to be underestimated by a factor of about two by the theory.

In fact, it may well be that the cases in which the theory does *not* agree with empirical measurements are really the most interesting. Consider again for a moment what our random graph models actually do. We have created random networks in which the degree distributions are the same as those for the real-world networks, but connections between vertices are otherwise random. If the real-world networks were also effectively random, then we would expect the predictions of our models to agree well with real-world measurements. The fact that in some cases the agreement is not perfect indicates lack of randomness, i.e., non-trivial structure, in these networks. In fact, there are some obvious possibilities for what this structure might be. We see for example that the clustering coefficient of the scientific collaboration networks is uniformly higher in real life than in the theory. This may indicate perhaps that scientists tend to introduce pairs of their collaborators to one another, encouraging new collaborations and hence producing higher clustering in the networks. There is some empirical evidence that indeed this is the case [30]. We see also that the typical number of a scientist's collaborators is lower than the number predicted by theory, which might arise because scientists are collaborating repeatedly with the same colleagues, rather than writing each paper with new and different coauthors. Thus, the discrepancy between theory and experiment may be highlighting real sociological phenomena in the networks studied. In a sense our random graph models of social networks are just providing a baseline against which real-world networks can be compared. Agreement between model and reality indicates that there is no statistical difference between the real-world network and an equivalent random network. Disagreement points to additional underlying processes, which may well be deserving of further investigation.

## VI. CONCLUSIONS

In this paper we have described and analysed a class of model networks which are generalizations of the much-studied random graph of Erdős and Rényi. We have applied these to the modeling of social networks. Our models allow for the fact that the degree distributions of real-world social networks are often highly skewed and quite different from the Poisson distribution of the Erdős–Rényi model. Many of the statistical properties of our networks turn out to be exactly solvable, once the degree distribution is specified. We have shown that there can be a phase transition with the parameters of the model at which a giant component of connected vertices forms, and given a formula for the position of this transition, as well as results for the size of the giant component and the average size of other smaller components. We can also calculate the average number of vertices a certain distance from a specified vertex in the network, and this result leads to a further expression for the typical distance between vertices in the network, which is found to increase only logarithmically with the size of the network. In addition we have generalized our theory to the case of bipartite random graphs, which serve as models for affiliation networks, and thus calculated such properties as clustering coefficients and average degree for affiliation networks.

We have compared the predictions of our models to a variety of real-world network data. Predictions for typical vertex–vertex distances, clustering coefficients, and typical vertex degree agree well with empirical data in some cases. In others they give results of the correct order of magnitude but differing from the empirical figures by a factor of two or more. We suggest that discrepancies of this sort indicate non-random social phenomena at work in the shaping of the network. Thus our models may provide a useful baseline for the study of real-world networks: if a comparison between a network and the equivalent random model reveals substantial disagreement, it strongly suggests that there are significant social forces at work in the network.

Finally, we point out that, although we have applied our models only to social networks in this paper, there is no reason why they should not be used in the study of other kinds of networks. Communication networks, transportation networks, distribution networks, metabolic networks, and food webs have all been studied recently using graph theoretic methods, and it would certainly be possible to apply the types of approaches outlined here to these systems. We have given one such application, to the World-Wide Web, in Ref. 15, and we hope that researchers studying other types of networks will find our methods of use also.

### Acknowledgments

The authors would like to thank Jerry Davis, Paul Ginsparg, Oleg Khovayko, David Lipman, and Grigoriy Starchenko for supplying data used in this study. This work was funded in part by the National Science Foundation, the Army Research Office, the Electric Power Research Institute, and Intel Corporation.

[1] S. Wasserman and K. Faust, *Social Network Analysis,* Cambridge University Press, Cambridge (1994).
[2] J. Scott, *Social Network Analysis: A Handbook,* 2nd Edition, Sage Publications, London (2000).
[3] S. H. Strogatz, “Exploring complex networks,” *Nature* **410**, 268–276 (2001).
[4] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.,* in press.
[5] R. Albert, H. Jeong, and A.-L. Barabási, “Error and attack tolerance of complex networks,” *Nature* **406**, 378–382 (2000).
[6] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, “Resilience of the Internet to random breakdowns,” *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
[7] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Network robustness and fragility: Percolation on random graphs,” *Phys. Rev. Lett.* **85**, 5468–5471 (2000).
[8] I. Pool and M. Kochen, “Contacts and influence,” *Social Networks* **1**, 1–48 (1978).
[9] S. Milgram, “The small world problem,” *Psychology Today* **2**, 60–67 (1967).
[10] J. Guare, *Six Degrees of Separation: A Play,* Vintage, New York (1990).
[11] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks”, *Nature* **393**, 440–442 (1998).
[12] R. Albert, H. Jeong, and A.-L. Barabási, “Diameter of the world-wide web,” *Nature* **401**, 130–131 (1999).
[13] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, “Classes of small-world networks,” *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).
[14] M. E. J. Newman, “The structure of scientific collaboration networks,” *Proc. Natl. Acad. Sci. USA* **98**, 409–415 (2001).
[15] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Phys. Rev. E* **64**, 026118 (2001).
[16] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science* **286**, 509–512 (1999).
[17] M. E. J. Newman, “A study of scientific collaboration networks: I. Network construction and fundamental results,” *Phys. Rev. E* **64**, 016131 (2001); “A study of scientific collaboration networks: II. Shortest paths, weighted networks, and centrality,” *Phys. Rev. E* **64**, 016132 (2001).
[18] P. Mariolis, *Social Science Quarterly* **56**, 425 (1975).
[19] G. F. Davis, M. Yoo, and W. E. Baker, “The small world of the corporate elite,” preprint, University of Michigan

Business School (2001).

[20] S. Redner, "How popular is your paper? An empirical study of the citation distribution," *Eur. Phys. J. B* **4**, 131–134 (1998)

[21] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *Comp. Comm. Rev.* **29**, 251–262 (1999).

[22] D. Fell and A. Wagner, "The small world of metabolism," *Nature Biotechnology* **18**, 1121–1122 (2000).

[23] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae* **6**, 290–297 (1959).

[24] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms* **6**, 161–179 (1995).

[25] M. Molloy and B. Reed, "The size of the giant component of a random graph with a given degree sequence," *Combinatorics, Probability and Computing* **7**, 295–305 (1998).

[26] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (2000).

[27] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South,* University of Chicago Press, Chicago (1941).

[28] J. Galaskiewicz and P. V. Marsden, *Social Science Research* **7**, 89 (1978).

[29] The Internet Movie Database can be found on the World-Wide Web at `http://www.imdb.com/`.

[30] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E* **64**, 025102 (2001).