Meeting Strangers and Friends of Friends: How Random are Social Networks?

Matthew O. Jackson and Brian W. Rogers *[0]

May 2004

Revised: October 27, 2006

Forthcoming, *American Economic Review*

Abstract

We present a dynamic model of network formation where nodes find other nodes with whom to form links in two ways: some are found uniformly at random, while others are found by searching locally through the current structure of the network (e.g., meeting friends of friends). This combination of meeting processes results in a spectrum of features exhibited by large social networks, including the presence of more high and low degree nodes than when links are formed independently at random, having low distances between nodes in the network, and having high clustering of links on a local level. We fit the model to data from six networks and impute the relative ratio of random to network-based meetings in link formation, which turns out to vary dramatically across applications. We show that as the random/network-based meeting ratio varies the resulting degree distributions can be ordered in the sense of stochastic dominance, which allows us to infer how the formation process affects average utility in the network.

JEL Classification Numbers: D85, A14, C71, C72.

Keywords: Networks, Network Formation, Power Laws, Scale-Free Networks, Small Worlds, Search.

Social network structures are important in determining outcomes in many settings. Situations where network structures play a central role include scientific collaborations among academics, joint research ventures among firms, political alliances, trade networks, the organization of intra-firm management, the sharing of information about job opportunities, and "peer-to-peer" computer systems for file sharing.[1] Given the prevalence of situations where social networks play a key role, it is important to understand network formation and how the process through which networks form affects their efficiency.

We examine a model of network formation that does several things. First, it leads to networks exhibiting characteristics that are common to large socially generated networks. Second, as parameters of the model are varied, the specific form of the emerging networks vary. This allows us to fit the formation model to data and estimate parameters corresponding to a range of observed networks. Third, we show how the efficiency of the emerging networks vary as the parameters of the model are varied.

A. Characteristics of Socially-Generated Networks

Before describing the model, let us provide some brief background. The key empirical regularities shared by socially generated networks can be summarized as follows.[2]

(i) The average distance (measured by shortest path length) between pairs of nodes in a social network tends to be small, and the maximum distance between any pair of nodes in a social network (called the diameter) is also small, where small is on the order of the log of the number of nodes or less.[3]

(ii) Clustering coefficients, which measure the tendency of linked nodes to have common neighbors, are larger in social networks compared to networks where links are generated by an independent random process.[4]

(iii) The distribution of degrees of the nodes in a network tends to exhibit "fat tails," so that there are more nodes with relatively high and low degrees, and fewer nodes with medium degrees, than one would find in a network where links are formed uniformly at random. In some cases, observed degree distributions have been claimed to be approximately "scale-free" or to follow a "power-law" distribution (e.g., see Reka Albert, Hawoong Jeong, and Barabási (1999), where the relative frequency of nodes with degree d is proportional to $d^{-\gamma}$ for some $\gamma > 1$.[5,6]

(iv) The degrees of linked nodes tend to be positively correlated, so that higher degree nodes are more likely to be linked to other higher degree nodes, and lower degree nodes are more likely to be linked to other lower degree nodes. This is referred to as (positive) assortativity.[7]

(v) The clustering among the neighbors of a given node, in at least some social networks, is inversely related to the node's degree. That is, the neighbors of a higher degree node are less likely to be linked to each other compared to the neighbors of a lower degree node.[8]

B. A Preview of the Model and Results

We now describe our model and preview why it exhibits these features. Nodes are born sequentially. When a new node is born, it meets some of the existing nodes through two processes. First, it meets some nodes uniformly at random. We refer to such meetings as "random meetings." Second, it then meets some of those nodes' neighbors in the current network. This means that the existing network structure influences how new links are formed. This second process is what we refer to as "network-based meetings." There is then a probability that the new node and any given node that it has met are compatible, and if they are then a link is formed.

Let us explain why this process exhibits features (i)-(v). Nodes with higher degree (more links) are more likely to be found through the network-based meeting process since more paths lead to them. This leads to an element of proportional growth in degree. Pure proportional growth would produce a degree distribution obeying a power law and, indeed, resulting degree distributions approximate a power law for high degree nodes. However, the fact that some links are formed through uniformly random meetings allows for richer distributions that can also exhibit non power-law characteristics. The temporal aspect of the model produces assortativity, since older nodes are both more likely to be linked to each other and more likely to have large degree.[9] Clustering results from the network-based meeting process since some nodes meet each other explicitly through common neighbors. The negative relationship between a node's degree and local clustering in its neighborhood comes from the fact that high-degree nodes attracted most of their neighbors via network-based meetings, and each of those neighbors then forms a relatively small number of connections to the node's neighbors. The small diameter results both from random meetings, which effectively form "bridges" between neighborhoods, and the prevalence of high-degree nodes or "hubs," which bring many nodes into close proximity.

By fitting this process to data from six different networks, we find widely different ratios of the role of random versus network-based meetings in link formation. Although the model is simple, it does a remarkable job of matching various network statistics across applications, and thus allows us to begin to trace differences in network characteristics back to differences in the formation process. For instance, we find that the relative roles of the random versus network-based meetings is more than eight times greater in a co-authorship network than in a world wide web application, and the formation process is almost uniformly random in two applications where links correspond to friendships.

The relative simplicity of the model also allows us to derive a tight relationship between parameters of the model and welfare differences in resulting networks. In particular, as the ratio of random to network-based meetings is varied, we can completely order the degree distribution in the sense of second order stochastic dominance. This is useful as it allows us to derive results about the efficiency of the resulting networks. For instance, if the utility derived by a node is a concave function of its degree, then second order stochastic dominance of the degree distribution implies that we can completely order the resulting networks in terms of total utility. These are the first results that we are aware of that tie variations in the stochastics behind network formation to variations in efficiency of resulting networks.

Formal modeling of network formation can be roughly split into two categories. One set examines efficiency and/or strategic formation of networks. These models use game-theoretic tools and lie more or less exclusively in the economics literature.[10] The other set is more mechanical, describing stochastic processes of network formation that exhibit some set of characteristics. This set has roots back in the early random graph literature, has overlap with the sociology literature, and has been very recently flourishing in the computer science and statistical physics literatures.[11] Our model has characteristics of both categories. The agents or nodes in our model are non-strategic. While the meeting process has many natural characteristics, it is in the tradition of the random graph literature in terms of being largely based on a mechanical (stochastic) process. On the other hand, we are able to tie the implications of the stochastic process back to the welfare of the agents, and thus deduce efficiency characteristics of the networks.

A variety of recent random graph models have been proposed to explain some of properties (i) to (v). For example, Watts (1999) and Watts and Steven H. Strogatz (1998) generate networks exhibiting the "small world" characteristics, (i) and (ii), by starting with a symmetric network and randomly rewiring some links.[12] Price (1976), Barabási and Albert (1999), and Colin Cooper and Alan Frieze (2003) have shown that networks with power law degree distributions, (iii), result if nodes form links through preferential attachment (i.e., new nodes link to existing nodes with probabilities proportional to the existing nodes' degrees).[13] Power-law distributions have also been shown to result if new nodes copy the links of a randomly identified node (Jon Kleinberg et al (1999) and Ravi Kumar et al (2000)), or if networks are designed to optimize tolerance (e.g., Jean Carlson and John Doyle (1999) and Alex Fabrikant et al (2004)).[14] A variation on preferential attachment where only some nodes are active at any time (Konstantin Klemm and Victor Eguíluz (2002a,b)) has been shown to also exhibit small world properties (i)-(ii). Some network models that grow over time have been shown to exhibit (iv) (e.g., Duncan S. Callaway et al (2001) and Paul L. Krapivsky and Sidney Redner (2002)).

While the above-mentioned models result in some of the empirical regularities of large social networks, none of them are not consistent with all of (i)-(v). Thus, those methods of generating networks cannot be the ones underlying most of the large networks that we actually observe.[15]

Our simple model exhibits all of the stylized facts by combining random meetings and network-based meetings in a natural way, and it is important to understand that the particular relationship between random meetings and network-based meetings is critical to obtaining our results. Models that mix random meetings and preferential attachment (e.g., vertex copying – see Jon Kleinberg et al (1999) and Kumar et al (2000)) or have nodes randomly decide to form their links one way or the other (see Pennock et al (2002)), cannot exhibit all of the features discussed here. In particular, we obtain the high clustering (ii), a diameter that is smaller than that of a random graph (i), and the negative clustering-degree relationship (v), precisely because each node has some chance of forming multiple links at random and *at the same time* forming other links within the neighborhoods found from that process.

The most similar models in structure to ours are by Alexei Vazquez (2003) and Pennock et al (2002). The model of Pennock et al (2002) adds random link formation to a preferential attachment model.[16] This leads to a degree distribution that is similar to the one from the model analyzed here. However, the fact that their link formation process is not network-based means that there is no limiting clustering in such a model, and other characteristics of their model are also quite different. Nevertheless, Pennock et al (2002) were the first to fit a family of degree distributions to network data sets, pointing out that observed networks do not actually match pure preferential attachment. The contribution of the empirical portion of our analysis relative to theirs is that we examine the model's fit along more dimensions than just the degree distribution. We also fit the clustering, diameter and assortativity that emerge from the network-based meetings approach, and find that the network-based model fits observed networks well on these other dimensions, in addition to fitting the degree distributions closely. Vazquez (2003) analyzes several models, including one where nodes enter by finding a starting node uniformly at random and then subsequently randomizing between following a random link from the last node visited, or hopping to a new node uniformly at random. While that model differs from ours in details, it is perhaps the closest from the literature in that it combines random meetings with some local search, and thus has a network-based meetings flavor to it. As such, his model has a degree distribution that differs from a pure power law, exhibits nonzero clustering and has a negative degree-clustering relationship. While Vazquez does not show his model to have all of the features mentioned above, a strong conjecture based on our analysis is that his model would have similar diameter and assortativity characteristics as we have found in our model. Vazquez does not fit his model to data or derive any stochastic dominance conclusions.

Perhaps the most useful innovation to come out of our analysis is showing that as one varies the parameters of the model (the average number of links formed and/or the ratio of uniformly random to network-based meetings), the resulting degree distributions can be ordered in the sense of first and second order stochastic dominance. This helps to tie the network formation process to outcomes and welfare, since when one can order degree distributions in this way, then one can order outcomes and utilities.[1]

I. The Network Formation Model

Networks are modeled as directed graphs. Given a finite set of agents or nodes N, a directed graph on N is an $N \times N$ matrix g where entry g_{ij} indicates whether a directed link exists from node i to node j. The obvious notation is that $g_{ij} = 1$ indicates the presence of a directed link and $g_{ij} = 0$ indicates the absence of a directed link.

Non-directed graphs are similarly defined, but with the restriction that $g_{ij} = g_{ji}$ for all nodes i and j.

For any node $i \in N$, let $d_i(g) = |\{j \in N \mid g_{ji} = 1\}|$ denote the in-degree of i. That is, $d_i(g)$ is the number of links that connect to i from other nodes.[17] Let $n_i(g) = \{j \in N \mid g_{ij} = 1\}$ denote i's neighborhood; i.e., the nodes that can be reached via links leaving node i.

The network is formed over time as follows. Link formation takes place at a countable set of dates $t \in \{1, 2, \ldots\}$. At each date t a new node is added to the population. Denote the node born at date t by its birthdate t. Let N_t denote the set of all nodes present at the end of time t. Denote by $g(t)$ the network consisting of the links formed on the nodes N_t by the end of time t.

Links are formed as follows.

Upon birth, node t identifies m_r nodes uniformly at random (without replacement) from N_{t-1}. We call these nodes t's "parent nodes." Node t forms a directed link to a given parent node if t's marginal utility from forming that link is positive. For now, assume that the

[1] See Jackson and Rogers (2004) for an analysis of how stochastic dominance can be used to order diffusion of disease and behavior through social networks.

marginal utility from forming a directed link is independently and identically distributed across pairs of nodes, regardless of the network structure. Let p_r denote the probability that any new node finds a parent node attractive to link to. In section we return to richer formulations of utility that allow for indirect benefits and externalities from the network structure.

The node t also meets other nodes in its parents' immediate neighborhoods. The new node t finds a total of m_n nodes through these network-based meetings. These nodes are picked uniformly at random without replacement from the union of the parents' neighborhoods (excluding all parents).[18] Let p_n denote the probability that node t obtains a positive marginal utility from linking to any given node found through each network-based meeting.

It would be natural to require that $p_r = p_n = p$, but we allow for the additional heterogeneity so that we can nest other models as special cases.[19]

In order for the process to be well-defined upon starting, begin with an initial network on a set of at least $m_r + m_n + 1$ nodes, where each node has at least $m_r + m_n$ neighbors.

In our model nodes have the same expected out-degree but will have different expected in-degrees depending on how long they have been alive. One could add heterogeneity in out-degree. This would add complexity without adding much insight to the analysis. Thus we focus primarily on in-degree.

This process forms a directed network. While some applications involve directed networks (e.g., a network of citations of articles by others), others are non-directed (e.g., a network of friends or acquaintances). We can adopt the process to deal with non-directed applications by working with the process as described, where the directed nature of the link indicates the node who "initiated" a meeting, but where every link is treated as if it represents a mutual acquaintance.

An expression for the probability that a given existing node i with in-degree $d_i(t)$ gets a new link in period $t+1$ is roughly

$$p_r \frac{m_r}{t} + p_n \left(\frac{m_r d_i(t)}{t} \right) \left(\frac{m_n}{m_r(p_r m_r + p_n m_n)} \right). \tag{1}$$

The first main expression in (1) is the probability that the node is found at random by the new node and then linked to. There are t existing nodes, and the new node picks m_r of them at random and links to them with a probability p_r. The second main expression in (1) is the probability that the node is found via a network-based meeting and is then linked to. This breaks into three parts. The term $\frac{m_r d_i(t)}{t}$, is the probability that some node with a link to i, say a node j, is chosen as a parent, so that i has the potential of being met in this way. The term $\frac{m_n}{m_r(p_r m_r + p_n m_n)}$ is then the probability that i is found given node j has been met randomly. The denominator, $m_r({}_r m_r + p_n m_n)$, is the expected size of the union of the parents' neighborhoods, and the numerator, m_n, is how many nodes are met out of that union. The factor p_n is the probability that a link is formed after the nodes have met via a network-based meeting.[20]

Letting $m = p_r m_r + p_n m_n$ be the expected number of links that a new node forms, we rewrite (1) as

$$\frac{p_r m_r}{t} + \frac{p_n m_n d_i(t)}{mt}. \tag{2}$$

A. A Mean-field Analysis of the Degree Distribution

The analysis of this model is complicated, especially given that the meetings depend on the structure of the network in place at each date. Thus, we use techniques that are common to analyzing such dynamic systems. In particular, we analyze a "mean-field" approximation to this system. This is a continuous time system where all actions happen deterministically at a rate proportional to the expected change. We then also run some simulations of the actual system and compare these to the predictions of the mean-field approximation, as well as checking the approximation in extreme cases where closed-form solutions can be derived.

Consider a process that evolves over time (continuously) where the in-degree of a given node i at time t changes deterministically according to

$$\frac{dd_i(t)}{dt} = \frac{p_r m_r}{t} + \frac{p_n m_n d_i(t)}{tm}. \tag{3}$$

Let each node t be born having in-degree counted as d_0. This can be set to 0, but allowing for other cases again allows us to nest other models.[21] We solve the differential equation (3) to find

$$d_i(t) = (d_0 + rm)\left(\frac{t}{i}\right)^{\frac{1}{1+r}} - rm,$$

where $r = \frac{p_r m_r}{p_n m_n}$ is the ratio of the number of links that are formed uniformly at random compared to through network-based meetings.[22] Depending on the application, the parameter r, might either capture the inherent randomness of the environment in terms of the meeting process, or might instead capture a search algorithm (like a variation on simulated annealing) that nodes choose to follow in meeting other nodes.

Using this expression for the in-degree of a node, we can derive the degree distribution by keeping track of how this varies across nodes.

Theorem 1 *The in-degree distribution of the above mean-field process has a cumulative distribution function of*

$$F_t(d) = 1 - \left(\frac{d_0 + rm}{d + rm}\right)^{1+r}, \tag{4}$$

for $d \geq d_0$ and each time t.

The proof of Theorem 1 follows from Lemma 1, which appears in the appendix.

As a check on the mean-field approximation, we compare the distribution function from Theorem 1 with degree distributions generated by simulations of the random process. The results are consistent for a variety of different parameters that we have checked.[23] Figure 1 shows a range of comparisons. The red curves are the predictions from (5) and the black dots depict empirical distributions from simulations.[24]

To get an idea of how the degree distribution resulting from the network-based meeting model relates to a scale-free distribution, we rewrite (4) as

$$\log(1 - F(d)) = \frac{m}{p_n m_n}\left[\log(d_0 + rm) - \log(d + rm)\right] \tag{5}$$

For large d relative to rm, (5) is roughly linear in $log(d)$, and thus approximates a scale-free distribution. *However, for small d, the expression does not approximate a scale-free distribution.* One can also see the effect of varying r on the degree distribution. As $r \to 0$, links are formed primarily via network-based meetings, which gives an advantage to higher degree nodes. This corresponds to "preferential attachment" (in the sense of Barabasi and Albert (1999)) and leads to a scale-free or power-law distribution.[25] At the other extreme where $r \to \infty$, the process is one of uniformly random link formation.[26] Figure 1 clearly

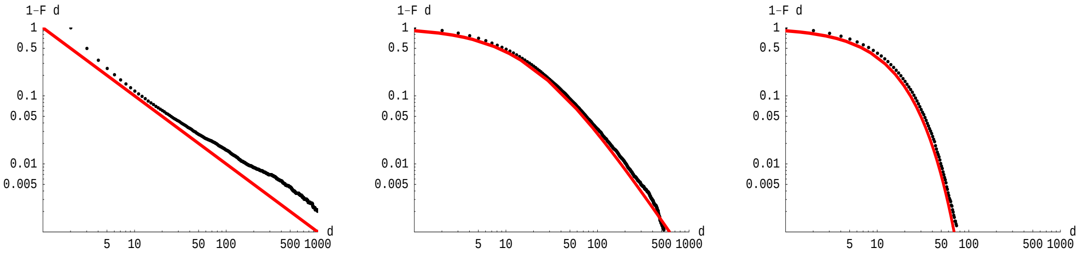

Figure 1. *The simulations cover* 25,000 *periods. The red curves are the predictions from (5) while the black dots depict empirical distributions from simulations.* **Left Panel:** *All links are formed through network-based meetings ($m_r = m_n = 10$, and $p_r = 0$, $p_n = 1$, starting each node with an in-degree of one to ensure that entering nodes can be found).* **Middle Panel:** *Equal numbers of uniformly random and network-based meetings, ($m_r = m_n = 10$, and $p_r = p_n = 1$).* **Right Panel:** *All meetings are uniformly at random ($m_r = m_n = 10$, $p_r = 1$, and $p_n = 0$).*

shouds how the model changes as r is varied. The left panel corresponds to the extreme of all network-based meetings, while the right panel is the other extreme of all random meetings, and the middle panel corresponds to a process with equal numbers of uniformly random and network-based meetings.

B. Clustering

We now analyze the clustering coefficients that emerge from networks formed under the model.[27] There are several ways to measure clustering, and we examine three common ones.

The first is a well-known measure from the sociology literature (e.g., see Stanley Wasserman and Katherine Faust (1994)) that examines the percentage of "transitive triples." This looks at situations where node i has a (directed) link to j, and j has a (directed) link to k, and then asks whether i has a (directed) link to k.[28] The percentage of times in a network that the answer is "yes" is the *fraction of transitive triples.* This fraction is represented as follows.

$$C^{TT}(g) = \frac{\sum_{i;j\neq i;k\neq j,i} g_{ij}g_{jk}g_{ik}}{\sum_{i;j\neq i;k\neq j,i} g_{ij}g_{jk}}.$$

While the above fraction of transitive triples is a standard measure, much of the empirical literature on large networks has considered variations of it where the directed nature of relationships is ignored, even though the relationships may indeed be asymmetric (e.g., links from one web pages to another; e.g. see Newman (2003)). That is, setting $\widehat{g}_{ij} = \max\left[g_{ij}, g_{ji}\right]$, we have an alternative measure of clustering where the directed nature of links is ignored and we only pay attention to whether there is some relationship between nodes. This total clustering measure is

$$C(g) = \frac{\sum_{i;j\neq i;k\neq j,i} \widehat{g}_{ij}\widehat{g}_{jk}\widehat{g}_{ik}}{\sum_{i;j\neq i;k\neq j,i} \widehat{g}_{ij}\widehat{g}_{jk}}.$$

These two measures clearly coincide when the network is not directed, but are different otherwise.

Another variation is similar to the total clustering coefficient $C(g)$ above, except that instead of considering the overall percentage of triples out of potential triples, one does this on a node-by-node basis and then averages across nodes. For example, this measure is used by Watts (1999), as well as many empirical studies (e.g., see Newman (2003)), and is calculated as follows.

$$C^{Avg}(g) = \frac{1}{n} \sum_i \frac{\sum_{j \neq i; k \neq j, i} \widehat{g}_{ij} \widehat{g}_{jk} \widehat{g}_{ik}}{\sum_{j \neq i; k \neq j, i} \widehat{g}_{ij} \widehat{g}_{ik}}.$$

These measures can differ significantly,[29] as we shall see.

In order to obtain closed-form expressions for these coefficients, we again use mean-field approximations. We also assume that m is an integer and that the process is such that if $\frac{p_r}{r} < 1$ then at most one link is formed in each parent's neighborhood, and otherwise we assume m_r and $\frac{p_r}{r} = \frac{p_n m_n}{m_r}$ to be positive integers and that exactly $\frac{p_r}{r}$ links are formed in each parent's neighborhood (where $\frac{p_r}{r} = \frac{p_n m_n}{m_r}$ represents the expected number of links formed through network-based meetings per parent node identified). This provides a tractable approximation to the more general process.

Theorem **2** *Under the above described mean-field approximation to the model:*

The fraction of transitive triples, C^{TT}, tends to

$$\begin{cases} \frac{p_r}{m(1+r)} & \textit{if } \frac{p_r}{r} \leq 1, \ \textit{and} \\ \frac{p_r(m-1)}{m(m-1)(1+r) - m(\frac{p_r}{r} - 1)} & \textit{if } \frac{p_r}{r} > 1. \end{cases}$$

Total clustering, $C(g)$, tends to

$$\begin{cases} 0 & \textit{if } r \leq 1, \ \textit{and} \\ \frac{6p_r}{(1+r)[(3m-2)(r-1)+2mr]} & \textit{if } r > 1. \end{cases}$$

Average clustering, $C^{Avg}(g)$, tends to

$$\int_{\max\{0,1-m\}}^{\infty} \left[\frac{(rm)^{r+1}(r+1)}{(d+rm)^{r+2}} \right] \left(\frac{m^2 C^{TT} \left(1 + \frac{2d(1+r)}{m}\right) - p_r d + rm \left[log\left(\frac{d}{rm} + 1\right) \right] \left(\frac{p_r}{r} + p_r - 2C^{TT} m(1+r) \right)}{(d+m)(d+m-1)/2} \right) dd$$

We remark that at the extremes of a uniformly random network ($p_n = 0$) and a pure preferential attachment network ($p_r = 0$), all three coefficients tend to 0. Those predictions are inconsistent with the empirical evidence suggesting that large decentralized networks exhibit significant clustering. For instance, Watts (1999) finds an average clustering coefficient of 0.79 in a network consisting of movie actors where links represent appearing in the same film, and Newman (2003) reports a total clustering coefficient of 0.20 for the same network. Networks of researchers where links indicate co-authorship have also been analyzed in various fields of study. Newman (2003) reports total clustering coefficients of 0.496 for networks of computer scientist, and 0.43 for physicists, while Jerrold W. Grossman (2000) reports a measure of 0.15 for a network of mathematicians. Several authors have also analyzed clustering in the world wide web. For instance, Lada A. Adamic (1999) reports an average clustering measure of 0.1078 on a portion of the web containing over 150,000 sites (which would have a predicted average clustering of only 0.00023 if the same number of links were formed uniformly at random over the same nodes).

In contrast, in variations of the model with both random and network-based meetings the fraction of transitive triples and the average clustering coefficient are positive.[30] The intuition for why these clustering coefficients are positive relies on the combination of the random and network-based meetings. A node is likely to link to two nodes who are linked to each other, precisely because they are linked to each other and one is found through a search of the other's neighbors. This is the critical feature that distinguishes the network-based meeting model from pure random graph models, preferential attachment models, and previous hybrid random graph and preferential attachment models where the preferential attachment and random attachment aspects are not tied to each other. As we shall in Section , the clustering coefficients generated by our model will fit those from observed networks fairly closely.

A final point is that the *total* clustering coefficient is nonzero only when $r > 1$, that is, only in cases where random meetings are more prevalent than network-based meetings. When $r < 1$, the predominance of links are formed through network-based meetings, and there is a high proportion of nodes with extremely high in-degrees. The impact on total clustering comes from the fact that nodes with huge in-degrees dominate the total clustering calculation since the calculation is effectively a degree-weighted average. The contrast with the average clustering calculation is due to the equal weighting of nodes in that calculation. Nodes with huge in-degrees have very low clustering rates since the many nodes that have connected to them are relatively unlikely to be linked to each other.[31]

C. Diameter

It is difficult even to derive bounds on the diameter of a random network, especially for models that are more complex than the uniformly random graphs studied by Erdös and Rényi (1960).

For some special cases we deduce limits on the diameter by piggy-backing on powerful results due to Bollobás and Oliver Riordan (2002). In particular, they show that a preferential attachment process where each node forms a single link (see also Bruce Reed (2003)) consists of a single component with diameter proportional to $\log(t)$ almost surely, while if more than one link is formed by each new node then the diameter is proportional to $\frac{\log(t)}{\log\log(t)}$. In our context, this covers the following special case, where we assume that at least two nodes met through the network meeting process come from different randomly located nodes' neighborhoods:

Theorem **3** *If $p_r = 0$, $p_n = 1$, $m_n \geq 2$, and $m_r \geq 2$, then the resulting network will consist of a single component with diameter proportional to $\frac{\log(t)}{\log\log(t)}$, almost surely.*[32]

The proof follows from Bollobás and Riordan (2002).[33]

The theorem implies that when at least two neighborhoods are searched, the diameter of the resulting network is much smaller than that of a uniformly random network. Results from simulations support the conjecture that this holds more generally, as we shall see shortly.

The constraint that $m_r \geq 2$ is critical to the result. Attachment to at least two independent neighborhoods allows a node to form a bridge between different existing neighborhoods of the network, thus reducing path lengths. For instance, if $m_r = 1$, and the existing network consisted of several separate components, a new node would end up just connecting to one of those components. The network would never become connected. We conjecture that increasing the parameters p_r and m_n and decreasing p_n (provided $m_r \geq 2$ and $p_n m_n \geq 2$) would not increase the diameter, as this leads to an increased number of links in the network, and at least as many links formed via network-based meetings as the minimal case now covered by the theorem. This is confirmed by following the heuristic test suggested on page 24 of Bollobás and Riordan (2002). However, changing the process changes the relative likelihood of connecting to different nodes, and once the parameters p_r and m_n are increased, the process is no longer covered by the Bollobás and Riordan (2002) approach and the diameter seems quite difficult to bound.

D. Assortativity

As Newman (2003) notes, a further feature distinguishing socially generated networks from other networks (e.g., uniformly random networks or those that are designed or controlled by some central actor) is that the degree of connected nodes tends to be positively correlated. This is referred to as assortativity since higher degree nodes have a tendency to be linked to each other, as in an assortive matching.

Generally, most models of networks where nodes are born over time will be assortative. [34] The basic intuition is that nodes with higher degree tend to be older nodes. As nodes must connect to pre-existing nodes, they always connect to nodes that are at least as old as they are. Thus, old nodes have at least some relative bias to be connected to other old nodes. In models where degrees grow with time, this means that nodes with higher degree are relatively more likely to be connected to each other, leading to a positive correlation among the degree of connected nodes in the network.

In fact, we can draw a stronger conclusion than simple correlation, as we now show. Let $F_i^t(d)$ denote the fraction of node i's in-degree at time t that comes from connections with other nodes that have in-degree d or less.[35] Thus, $1 - F_i^t(d)$ represents the fraction of node i's in-degree that comes from connections with other nodes that have in-degree greater than d.

Theorem 4 *Under the mean-field approximation to the model (with nontrivial network-based meetings, so that $p_n m_n > 0$), if $d_i(t) > d_j(t)$, then $1 - F_i^t(d) > 1 - F_j^t(d)$ for all $d < d_i(t)$.*[36]

This result provides a stronger relationship than positive correlation. The distribution of the degrees of the neighbors of a relatively higher degree node first order stochastically dominates the corresponding distribution for a node with a lower degree.

E. Negative Clustering-Degree Relationship

Beyond looking at total and average clustering coefficients, we can examine how the clustering varies across different nodes' neighborhoods. Considering a node i, its neighborhood's clustering is the fraction of pairs of i's neighbors that are linked to each other. A network has a *negative clustering-degree* relationship if this local clustering coefficient is smaller for higher degree nodes, on average. This has been found, for instance, in co-authorship networks (e.g., see Goyal van der Leij, and Moraga-González (2003)) among others.

The model exhibits a negative clustering-degree relationship. Nodes with relatively higher degrees tend to have gained more links from network-based meetings. Since each node that links to i links to only a fixed number of other nodes on average, as i's degree grows, each new neighbor links to a smaller fraction of i's neighbors, and clustering decreases.[37]

This negative clustering-degree relationship is the reason why the total clustering coefficient $C(g)$ tends to zero while the average clustering coefficient $C^{Avg}(g)$ remains bounded away from zero when $r < 1$ (i.e., the role of random meetings is smaller than that of network-based meetings). Nodes with the largest degree have the smallest clustering in their neighborhoods (tending to 0 as degree grows), and receive a greater weight in the calculation of the total clustering coefficient.

Ideally we would like to show that $C(d)$, the clustering coefficient for a node with in-degree d, is a strictly decreasing function of degree d. While we believe this to be true, we do not have a proof. Nonetheless, we can prove a weaker version of this statement. In particular, we show that if two nodes have high enough degrees and have a large enough difference in their degrees, then the node with higher degree has a smaller local clustering coefficient. We make this precise in the following way.

Theorem **5** *Under the mean-field approximation to the model (with nontrivial meetings of both types so that $p_r m_r > 0$ and $p_n m_n > 0$), there exists $\bar{d} > 0$ such that for all $d > \bar{d}$ there exists $D > 0$ so that for all $d' > d + D$, $C(d) > C(d')$.*

II. Fitting the Model to Data

To demonstrate the power and flexibility of the model and to illustrate how it can identify elements of link formation, we calibrate it with data from widely varied applications.

We fit the model to six distinct data sets: the links among web sites on the Notre Dame www, the network of co-authorship relations among economists publishing in journals listed by Econlit in the 1990's, a citation network of research articles stemming from Milgram's 1960 paper (all papers either reference Milgram (1960) or contain the phrase "small world" in the title), a friendship network among 67 prison inmates in the 1950s, a network of ham radio calls during a one month period, and finally a network of romantic relationships among high school students.[38] We show that the characteristics predicted by the model closely match the observed characteristics of these networks.

We fit the model as follows. First, we directly calculate average in-degree to obtain m. Fixing m and setting $d_0 = 0$, the only free parameter remaining in the degree distribution F from (4) is r. As r enters F in a form $(d + rm)^{-(1+r)}$, we use an iterative least squares procedure where we start with an initial guess for r, say r_0 and plug this in to get an expression of the form $(d + r_0 m)^{-(1+r)}$.[39] We then regress $\ln(1 - F(d))$ on $\ln(d + r_0 m)$ to estimate $-(1 + r)$, and get an estimate r_1. We iterate this process until we find a fixed point r^*.[40] Next, we fit the clustering. We constrain $p_n = p_r = p$, as this seems a reasonable starting assumption and it eliminates a degree of freedom.[41] Then, using our expressions for clustering we can estimate p (which is not identified by the degree distribution).[42] This provides estimates of the parameters of the model, and allows us to compare the predicted degree distribution and clustering measures with the observed ones.

Beyond the degree distribution and clustering measure, we can also compare the model's predicted diameter with the observed diameter in each network, and similarly for the assortativity and degree-clustering relationships. Here, though, we need to resort to simulation. For instance, we only have an order of magnitude calculation of diameter from Theorem 3. This is not so useful empirically as even in the www data set, where the number of nodes is almost $T = 326,000$, one finds that $\ln(T) = 12.7$ and $\frac{\ln(T)}{\ln\ln(T)} = 5.0$. Thus, simply knowing an order of magnitude calculation does not even differentiate between a uniformly random network and one coming from the model. Even with the simulations, it is hard to calculate diameter directly. For instance, with the www data, the size of the network is larger than we can calculate using our program which can only take 100000 nodes, so we ran simulations based on $T = 10000$, 50000, and 100000, where we find lower bounds of 4, 4, and 5, respectively.

The reported figure is then obtained by extrapolation. The reason we obtain bounds rather than a point estimate is that for any given network, finding the precise diameter involves exponentially many calculations (in the number of nodes), which is impossible to compute in large networks. Thus we obtain lower bounds by starting from a node with maximal degree in the largest component, and then estimating the maximal shortest path from this node to any other, which provides a lower bound on the diameter. Doubling this estimate provides an upper bound. The variance on these bounds across simulations is remarkably small, generally varying by at most one.

The co-authorship and ham radio networks are non-directed.[43] Given the directed nature of our process, to fit the non-directed cases we adapt our model by keeping track of which nodes initiate the links, and let p represent the probability that *both* nodes find a link worthwhile.[44] For instance, in the co-authorship data, given that there are roughly two links per researcher, on average each researcher initiates one link and receives one link initiated by another. Accordingly, we set $m = .84$ and then estimate r as in the directed network case, inferring the in-degree distribution by taking in-degree to be one half of overall degree.

The following table reports on the six different networks and fits.[45]

The fits suggest, for example, that the role of network-based meetings is much more prevalent in the formation process of the www network than for the co-author network. Specifically, the random to network-based meetings ratio is approximately .57 in the www data and 4.7 in the economics co-author network. Thus network-based meetings are more than eight times more prevalent in the www network formation process than in the formation of the co-author network. We also see the two most purely social networks, the Prison and High School Romance networks, are almost uniformly random; in contrast, the world wide web and citations networks are more influenced by network-based meetings. The network formation processes in different settings can differ quite substantially and in a well-quantifiable manner.

Clustering measures are missing from the High School Romance data set since this is predominately a heterosexual network and so clustering is absent.[46] The diameter is also missing for the Romance data set, but we can still compute a prediction from the model from parameters satisfying our estimates for m and r. The www only has average path

Table 1: Parameter Estimates Across Applications

Data Set:	WWW	Citations	Co-author	Ham Radio	Prison	High School Romance
Number of Nodes:	325729	396	81217	44	67	572
Avg. In-Degree: m	4.6	5.0	.84	3.5	2.7	.83
r from Fit	0.57	0.63	4.7	5.0	∞	∞
p from Fit	.36	.27	.10	1	1	-
R^2 of Fit	.97	.98	.99	.94	.94	.99
Avg. Clustering Data	.11	.07	.16	.47	.31	-
Avg. Clustering Fit	.11	.07	.16	.22	.10	-
Diameter Data	11.3 (avg)	4	26	5	7	-
Diameter Fit	(6,12)	(4,8)	(19,38)	(4,8)	(5,10)	(12,24)

length reported (see Newman (2003)) rather than diameter, and so we report that. The www clustering figure of 0.11 is as reported in Adamic (1999) for her different www data set; the figure is not available for the Albert, Jeong, and Barabási (1999) data.[47]

Finally, we comment on the assortativity and clustering-degree relationships, which we have covered in the theoretical results. We have these measures only for the Prison and Ham radio data sets, as they are the only applications for which we have the complete networks.

For the prison network, the correlation between a node's in-degree and the average in-degree of its neighbors in the actual network is .58 (with a P-value of 0). In simulations based on our estimated parameters the correlation ranges from .69 to .83 (with a mean of .75 and a standard deviation of .07 across simulations).[48] For the Ham network, the correlation between a node's in-degree and the average in-degree of its neighbors is -.26, which is of the opposite sign from that predicted by the model (but is not statistically significant, with a P-value of .09). In simulations we find a range of correlations from -.35 to .22, with an average correlation of .02 and a standard deviation of the correlation across the simulations of .18. Given the small network size, there is substantial variation across simulations with a range that includes the observed data, and this variation might account for the observed negative relationship.

Estimating the relationship between clustering and degree, we find the predicted negative correlation in both the prison and Ham data sets, but neither are significant. For the prison, the correlation is -.05 (with a P-value of .70), while it is -.27 (with a P-value of .08) in the Ham data set. When we run simulations based on the fitted parameters for the Ham data set we find correlations ranging from -.08 to -.27, with and average of -.18 and a standard deviation of .10. Again, the small size of the network leads to substantial variation with a range that includes the observed $-.27$. When we run simulations based on the fitted values for the prison data, we find values ranging from .03 to .42, with an average of .17 and a standard deviation of .13. Here the simulations lead to correlations between degree and clustering that are positive. When examining the data the explanation becomes apparent. The average clustering is fairly low (.10), which is expected in a nearly uniformly random network. Then, much of the positive correlation is driven by the fact that a few of the highest degree nodes have above-average clustering. The highest degree nodes also tend to be the first born nodes, and when they form their links there are so few nodes to form links with that they naturally cluster among themselves. Later-born nodes have fewer links and clustering that is closer to 0. Recall that our result on the negative clustering-degree relationship is a limiting result and presumes that $p_n m_n > 0$. In a uniformly random network formation process, the clustering will tend to 0, but early born nodes will have non-trivial clustering when the network is still small, and so finite samples tend to exhibit a slightly positive relationship.

The following figures show the fits of the degree distributions.

In reference to the degree distribution of the www: even though the fitted model is clearly not scale-free ($r = .57$ implies that roughly one third of links are formed uniformly at random), it looks quite linear in the above plot (Figure , top left panel).[49]

III. Efficiency and Network Structure

Given that networks exhibit different average degree and ratios of uniformly random to network-based meetings, it is important to to understand the implications of the network formation process on the operation of a network. That is, we would like to know whether a high or low ratio of random to network-based meetings is a "good" or "bad" thing in terms of the functioning of a network. Based on the model, we can say quite a bit about this. A helpful result is that as we vary r, the resulting degree distributions are ordered in terms of strict second order stochastic dominance.

Theorem **6** *Consider the distribution function described in (4) with any fixed $m > 0$ and set $d_0 = 0$.*[50] *If $r' > r$, then F' strictly second order stochastic dominates F, where F' and F are the distribution functions corresponding to r' and r, respectively.*

Theorem 6 has powerful implications. One direct corollary is that if agents' utilities can be expressed as a concave function of their degree, then we can order the total utility of a network with respect to r. Concavity of utility in degree implies that nodes realize diminishing marginal utilities to additional links.

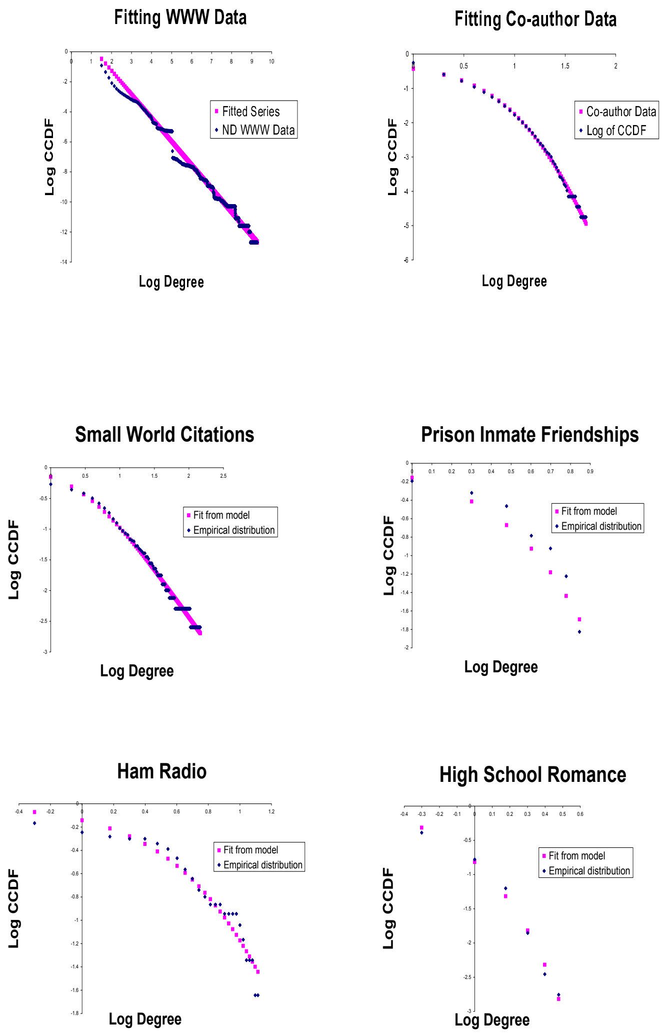

Figure 2. *Pink: Fit of complementary cdf from our model; Blue: complementary cdf from the data; (top left) Notre Dame www data set, (top right) Economics co-author data set, (middle left) Small Worlds citation network, (middle right) Prison friendship network, (bottom left) Ham-radio network, (bottom right) High school romance network.*

Corollary 1 *Suppose that the expected utility of a node in a network is a concave function of the node's degree, and the network's degree distribution is described by (4) with $d_0 = 0$. Then for any given m, if $r' > r$, then the average expected utility of agents in the network with r' is weakly higher than that under r, with a strict ranking if the expected utility function is strictly concave in degree.*[51]

In Theorem 6 we vary r while holding m fixed. Doing the reverse, we find that higher values of m correspond to degree distributions that *first* order stochastic dominate degree distributions corresponding to lower values of m. This has obvious implications for situations where utility is increasing in degree.

Theorem 7 *Consider the distribution function described in (4) and any fixed $r > 0$. If $m' > m$, then F' strictly first order stochastic dominates F, where F' and F are the distribution functions corresponding to m' and m, respectively.*

We omit the proof, as it is easily verified by noting that $1 - \left(\frac{d_0+rm}{d+rm}\right)^{1+r}$ is decreasing in m for any $d > d_0$.

While there are contexts where the expected utility of a node can be described, or at least approximated, by a concave and/or increasing function of the node's degree, there are also some important contexts where we cannot express welfare in this manner. For instance, in some cases the degrees of a node's neighbors are important in determining the given node's utility. While the above theorems do not apply directly to such cases, they are still useful to the extent that one can then deduce things about the ordering of outcomes from the ordering of degree distributions (e.g., see Jackson and Rogers (2004).

IV. Concluding Discussion and Extensions

We have presented a model of network formation that exhibits features matching observed socially generated networks. Fitting the model to data shows a wide range in terms of the relative ratio of uniformly random versus network-based meetings in the process of link formation, and also shows close fits of the model in terms of a variety of network properties.

We also show that the degree distributions coming from the model are ordered in the sense of first and second order stochastic dominance as parameters are varied. This can be useful in further applications, since if there is a well-structured relationship between degree and payoffs, then we can relate the network formation process to total societal utility.

The model and analysis suggest a pressing and interesting question regarding what accounts for the differences in the network formation process across applications. Why is it that the more purely social friendship networks appear to be governed largely through random meetings, while the world wide web and citation networks involve much more network-based link formation? Is there something systematic that can be said about network setting and the process through which nodes meet, and when it is that meetings should be more or less tied to existing network structure?

The power of the model and analysis comes at some cost. First, our approach uses techniques from mean-field analysis, which are commonly used in the study of complex dynamic systems. Relatively little is known about the circumstances where such analyses result in accurate approximations. We have checked that simulations of the model result in characteristics consistent with the approximations, but there are no results proving that the approximations are tight. Deriving such results seems to be a formidable challenge, even under severe restrictions on parameters. Second, our approach is largely mechanical in terms of the specification of the process, with little modeling of the reasons why links are formed in this way. Nonetheless, the model fits data remarkably well, and we can derive implications of the process for welfare. However, it is of interest to delve deeper into the micro details of link formation. With that in mind, let us discuss how the model extends along several dimensions.

A. Degree-Dependent Utility and Externalities

We have assumed that the utility obtained by a node from connecting to another is randomly drawn and independent of the rest of the structure of the network. In many contexts, the utility would instead be network- and degree-dependent. It might be that nodes benefit from indirect connections, and thus might be more willing to link to nodes that have larger degrees.[52] Alternatively, there might be correlation in the valuations across different nodes, and so higher degree might be related to a higher expected valuation for other nodes.

A simple variation on the model where the utility from attaching to a given node is proportional to its degree is as follows. Let the marginal utility obtained from linking to a node j be

$$u_{ij}d_j - c,$$

where u_{ij} is a random factor, say distributed uniformly on an interval $[0, u]$, and where c $(0 < c < u)$ is a cost parameter. Here u_{ij} might capture the compatibility of node i with j and the nodes that j has chosen to connect to. Then the probability of linking to a given node that has been identified via the network-based meeting process is proportional to $1 - \frac{c}{ud_j}$ (noting that it is necessary that $d_j \geq 1$ for j to have been met in this way).

Let us see how this would affect the degree distribution. For simplicity, suppose that parent nodes are attached to with certainty and that the utility calculation is only relevant for nodes identified through network-based meetings. We then obtain a mean-field process governed by

$$\frac{dd_i(t)}{dt} = \frac{m_r}{t} + \left(1 - \frac{c}{ud_i(t)}\right)\frac{m_n d_i(t)}{tm_t},$$

or

$$\frac{dd_i(t)}{dt} = \frac{m_r - \frac{m_n c}{m_t u}}{t} + \frac{m_n d_i(t)}{tm_t},$$

where $m_t = m_t(d_i)$ is the expected neighborhood size of a random parent node identified at time t, which is correlated with $d_i(t)$. In the limit (as t grows), m_t approaches a constant (it is growing and bounded above holding i constant, regardless of d_i), and so for this heuristic approximation, let m be that limit. By Lemma 1 (see the appendix),

$$1 - F_t(d) = \left(\frac{d_0 + rm - \frac{c}{u}}{d + rm - \frac{c}{u}}\right)^{m/m_n}.$$

While the parameters have changed, the basic expression is similar to (4) above. Essentially, this utility calculation tilts things more towards attaching to higher degree nodes – so that the degree distribution is closer to being scale-free over higher degrees, resulting in fatter tails to the distribution for any given r. Future research should investigate how other utility formulations impact the network formation process.

B. Non-directed Networks

In fitting the model to data, we explained how the analysis of the directed network model extends to the case of non-directed networks by keeping track of who initiates a link. That is restrictive in that the network-based meetings only occur through the links that a parent node initiated. One could also consider a variation where network-based meetings occur through all of a parent's links, rather than just the ones they initiated. This leads to some complications as now parent nodes have degrees that are correlated with their age and that of their connections.[53] This mitigates the degree-dependence of network-based meetings, as large degree nodes are more likely to be connected to each other, simply due to their age. As a result, since a given node is more likely to be found via neighbors who have fewer links, this counter-acts the benefit of having a high degree.

C. Larger Neighborhoods for Network-Based Meetings

Suppose that we alter the model so that network-based meetings extend uniformly over more extended (directed) neighborhoods of the parent node, instead of just to nodes directly connected to the parent node. For instance, meeting nodes within a distance of k to the parent node. This would lower clustering coefficients, but not to 0. More generally, the last expression in the probability of attachment for a given node in (1) would change, but other than that the calculations remain the same.

D. Exponential Population Growth

The model has a single node born at each point in time. Exponential growth in the number of nodes over time can also be incorporated, and is consistent in many applications. To see how, let us examine an extension of the model where the number of entering nodes at each date is proportional to population size, as in any naturally growing society. Let the number of new nodes entering at time t be gn_t, where n_t is the number of nodes at time t and $g > 0$ is a growth rate. The mean-field equation for degree evolution is then

$$\frac{dd_i(t)}{dt} = \frac{gn_t p_r m_r}{n_t} + \frac{gn_t p_n m_n d_i(t)}{n_t m}.$$

As shown in Lemma 3 in the appendix, this results in a distribution described by

$$F(d) = 1 - \left(\frac{d_0 + rm}{d + rm} \right)^{(1+r)\log(1+g)/g},$$

for $d \geq d_0$. This has a similar structure to that for the case of linear growth, except for the exponent.

E. Out-degree, Network-Based Meetings Initiated by Existing Nodes, and Death

One dimension along which our model is clearly too restrictive is that all nodes have (roughly) the same out-degree, except for randomness through the realization of which links form. One easy extension would be to allow for heterogeneous m_r, m_n, p_r, and p_n across nodes. More generally, in many network applications, nodes continue to add and delete links over time, and nodes leave networks. The main complications in such extensions is that the out-degree of nodes changes over time, which complicates some of the expressions in the mean-field analysis. This is clearly worthy of further analysis.

F. Other Applications

Although we have interpreted our model as a network-based meeting model of network growth, there are broader applications for which the model is of interest. There are many contexts where power laws have been observed, including city size.[54]

The data in the Figure 3 represent the population sizes of all counties in the US.[55] While the upper tail of the distribution is roughly linear, consistent with the fact that the literature has claimed scale-free distributions for (large) city populations.[56] It is important to remark, however, that the graph has a noticeable bend to it.[57] This point is emphasized by Eeckhout (2004) in the context of city sizes.

To briefly see how our model might relate to county size, note that county sizes are determined by the housing choices of individuals who are born into society and must choose where to live. Some individuals' choices are made randomly (at least to an outside observer) while others are determined by a preference to live close to friends or family, or close to a job location. This results in a similar random vs network-based choice process that would have the features of the process that we have analyzed.

References

[1] Adamic, Lada A. (1999) “The Small World Web,” *Proceedings of the ECDL* vol 1696 of Lecture Notes in CS, 443-454.

[2] Albert, Reka, and Albert-Laszlo Barabási (2002), “Statistical mechanics of complex networks,” *Reviews of Modern Physics,* **74**: 47-97.

[3] Albert, R., Hawoong Jeong, and A. Barabási (1999), “Diameter of the World Wide Web,” *Nature*, 401, 9 Sept., 130-131.

[4] Alderson, David (2004) ”Understanding Internet Robustness and its Implications for Modeling Complex Networks,” presentation.

[5] Barabási, A. (2002), *Linked*, Perseus Publishing: Cambridge, MA.

[6] Barabási A. and R. Albert (1999), “Emergence of scaling in random networks,” *Science*, **286**: 509-512.

[7] Barabási, A., R. Albert, and H. Jeong (1999), “Mean-field theory for scale-free random networks,” *Physica A* **272**: 173-187.

[8] Barabási, A., R. Albert, and H. Jeong (2000), “Scale-Free Characteristics of random networks: the topology of the world-wide web,” *Physica A* **281**: 69-77.

[9] Bearman, Peter, James Moody, and Katherine Stovel (2004), “Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks,” Manuscript, University of Chicago.

[10] Bollobás, Bela, and Oliver Riordan (2002), “The diameter of a scale-free random graph,”, Manuscript, to appear.

[11] Bollobás, B., O. Riordan, Joel Spencer, and Gabor Tusnady (2002), "The degree sequence of a scale-free random graph process," Manuscript.

[12] Callaway, Duncan S., John E. Hopcroft, Jon M. Kleinberg, Mark E.J. Newman, and Steven H. Strogatz (2001) "Are randomly grown graphs really random?" *Phys. Rev. E.*, 64, 041902.

[13] Nicolas Carayol, N. and Pascale Roux (2003) " 'Collective Innovation' in a Model of Network Formation with Preferential Meeting," mimeo: Université Louis Pasteur and Université de Toulouse I.

[14] Carlson, Jean, and John Doyle (1999), "Highly optimized tolerance: a mechanism for power laws in designed systems. *Physical Review E,* **60(2)**: 1412-1427.

[15] Cooper, Colin and Alan Frieze (2003) "A General Model of Web Graphs," preprint: Department of Computer Science, King's College, University of London.

[16] Dorogovtsev, Sergey N. and José F.F. Mendes (2001) "Scaling Properties of Scale-Free Evolving Networks: Continuous Approach," *Physical Review Letters*, 63: 056125.

[17] Eeckhout, Jaan (2004) "Gibrat's Law for (All) Cities," *American Economic Review*, vol. 94, no. 5, 1429-1451, Dec.

[18] Erdös, Paul and Alfréd Rényi (1960) "On the Evolution of Random Graphs," Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 5, 17-61.

[19] Fabrikant, Alex, Elias Koutsoupias, and Christos H. Papadimitriou (2002), "Heuristically Optimized Tradeoffs: A new paradigm for power laws in the Internet," *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming.*

[20] Fabrikant, Alex, Ankur Luthra, Elitza Maneva, Christos H. Papadimitriou, and Scott Shenker (2004) "On a Network Creation Game," preprint: U.C. Berkeley.

[21] Faloutsos, Michalis, Petros Faloutsos, and Christos Faloutsos (2004) "On Power-Law Relationships of the Internet Topology," preprint: U.C. Riverside.

[22] Gabaix, Xavier (1999) "Zipf's law for Cities: An Explanation," *Quarterly Journal of Economics*, August, pp 739-767.

[23] Galeotti, Andrea, Sanjeev Goyal, and Jurjen Kamphorst (2004) "Network formation with heterogeneous players," preprint: Tinbergen Institute.

[24] http://www.garfield.library.upenn.edu/histcomp/index.html.

[25] Gell-Mann, Murray (1994) *The Quark and the Jaguar*, Freeman: NY.

[26] Goyal, S., Marco van der Leij, and José-Luis Moraga-González (2003). "Economics: an emerging small world?," forthcoming: *Journal of Political Economy.*

[27] Grossman, Jerrold W. (2000) "The Evolution of the Mathematical Research Collaboration Graph," Proceedings of 33rd Southeastern Conference on Combinatorics (Congressus Numerantium, Vol. 158, 2002, pp. 201-212).

[28] Jackson, Matthew O. (2004) "A Survey of Models of Network Formation: Stability and Efficiency," in *Group Formation in Economics; Networks, Clubs and Coalitions* , edited by Gabrielle Demange and Myrna Wooders, Cambridge University Press: Cambridge U.K.

[29] Jackson, M.O. (2006) "The Economics of Social Networks," in *Proceedings of the 9th World Congress of the Econometric Society*, edited by Richard Blundell, Whitney Newey, and Torsten Persson, Cambridge University Press.

[30] Jackson, M.O. and Brian W. Rogers (2004) "Relating Network Structures to Diffusion Properties through Stochastic Dominance," forthcoming: *Advances in Theoretical Economics.*

[31] Jackson, M.O. and B. W. Rogers (2005) "The Economics of Small Worlds," *Journal of the European Economic Association – Papers and Proceedings)*, 3:(2-3), 617-627.

[32] Jackson, M.O. and Asher Wolinsky (1996) "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, Vol. 71, No. 1, pp 44–74.

[33] Kesten, Harry (1973), "Random difference equations and renewal theory for products of random matrices," *Acta Mathematica,* **CXXXI**: 207-248.

[34] Killworth, Peter D. and H. Russell Bernard (1976) "Informant accuracy in social network data," *Human Organization*, 35: 269-286.

[35] Kleinberg, Jon M., S. Ravi Kumar, Pranhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, (1999) "The Web as a graph: Measurements, models and methods," in Proceedings of the International Conference on Combinatorics and Computing, *Lecture Notes in Computer Science*, 1627, 1-18.

[36] Klemm, Konstantin and Victor M. Eguíluz (2002a) "Growing Scale-Free Networks with Small World Behavior," Physical Review E, vol 65(3), 036123,

[37] Klemm, K. and V.M. Eguíluz (2002b) "Highly Clustered Scale-Free Networks," Physical Review E, vol 65(5), 057102.

[38] Krapivsky, Paul L. and Sidney Redner (2002) "A Statistical Physics Perspective on Web Growth," *Computer Networks*, Vol. 39 No. 3, 261-276.

[39] Kumar, R., P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, Eli Upfal (2000) "Stochastic Models for the Web Graph" *FOCS 2000.*

[40] Levene, Mark, Trevor I. Fenner, George Loizou, and Richard Wheeldon (2002) "A Stochastic Model for the Evolution of the Web," *Computer Networks*, vol 39: 277-287.

[41] Li, Lun, David Alderson, Walter Willinger, John Doyle, Reiko Tanaka, and Steven Low (2004) "A First Principles Approach to Understanding the Internet's Router Technology," *Proc. Sigcomm*, ACM.

[42] MacRae, Duncan (1960) "Direct factor analysis of sociometric data," *Sociometry*, 23, 360-371.

[43] Milgram, Stanley (1967), "The small-world problem," *Psychology Today,* **2**: 60-67.

[44] Mitzenmacher, Michael (2004) "A Brief History of Generative Models for Power Law and Lognormal Distributions.", *Internet Mathematics*, **1(2)**: 226-251.

[45] Newman, Mark (2003), "The structure and function of complex networks," *SIAM Review*,**45**, 167-256.

[46] Newman, Mark (2004) "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, **101**: 5200-5205.

[47] Pareto, Vifredo (1896) "Cours d'Economie Politique." Droz, Geneva Switzerland.

[48] Pennock, David M., Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles (2002) "Winners don't take all: Characterizing the competition for links on the web," *PNAS*, 99:8, pp. 5207-5211.

[49] Price, Derek J. deSolla (1965) "Networks of Scientific Papers," *Science*, 149: 510-515.

[50] Price, D.J.S. (1976) "A General Theory of Bibliometric and Other Cumulative Advantage Processes," J. Amer. Soc. Inform. Sci. 27: 292-306.

[51] Reed, Bruce (2003) "The Height of a Random Binary Search Tree," *Journal of the ACM*, 50:3, pp 306-332.

[52] Rothschild, Michael and Joseph Stiglitz (1970) "Increasing Risk: I. A Definition,"*Journal of Economic Theory* 2: 225-243.

[53] Simmel, Georg (1908), "Sociology: Investigations on the Forms of Sociation," Duncker & Humblot, Berlin Germany.

[54] Simon, Herbert (1955), "On a class of skew distribution functions," *Biometrika,* **42(3,4)**: 425-440.

[55] Vázquez, Alexei (2003) "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations," *Physical Review E*, **67(5)**, 056104.

[56] Wasserman, Stanley and Faust, Katherine (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press.

[57] Watts, Duncan (1999), "Small Worlds," Princeton University Press.

[58] Watts, D. and S.H. Strogatz (1998), "Collective dynamics of 'small-world' networks," *Nature,* **393**: 440-442.

[59] Yule, G. Udny (1925), "A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis," *F.R.S. Philosophical Transactions of the Royal Society of London (Series B),* **213**: 21-87.

[60] Zipf, Geroge K. (1949) *Human Behavior and the Principle of Least Effort*, Addison-Wesley: Cambridge, MA.

V. Appendix

Lemma **1** *Consider a process where the degree of a node born at time i has initial degree d_0 and evolves according to*

$$\frac{dd_i(t)}{dt} = \frac{ad_i(t)}{t} + \frac{b}{t} + c.$$

If $a > 0$ and either $c = 0$ or $a \neq 1$, then the complementary cdf

$$1 - F_t(d) = \left(\frac{d_0 + \frac{b}{a} - \frac{ct}{1-a}}{d + \frac{b}{a} - \frac{ct}{1-a}} \right)^{1/a}. \tag{A1}$$

If $a = 0$ and $c = 0$, then (A3) has solution

$$1 - F_t(d) = e^{\frac{d_0 - d}{b}}. \tag{A2}$$

The proof of Lemma 1 uses the following lemma whose standard proof is omitted.

Lemma **2** *Consider a differential equation of the form*

$$\frac{dd(t)}{dt} = \frac{ad(t)}{t} + \frac{b}{t} + c, \tag{A3}$$

with initial condition $d(i) = d_0$ (where $i < t$). If $a > 0$ and either $c = 0$ or $a \neq 1$, then (A3) has solution

$$d(t) = \left(d_0 + \frac{b}{a} - \frac{ct}{1-a} \right) \left(\frac{t}{i} \right)^a - \frac{b}{a} + \frac{ct}{1-a}.$$

If $a = 0$, then (A3) has solution

$$d(t) = d_0 + b \log \left(\frac{t}{i} \right) + c(t - i).$$

Proof of Lemma 1: By Lemma 2 we can write

$$d_i(t) = \left(d_0 + \frac{b}{a} - \frac{ct}{1-a}\right)\left(\frac{t}{i}\right)^a - \frac{b}{a} + \frac{ct}{1-a}. \tag{A4}$$

if $a > 0$ and either $c = 0$ or $a \neq 1$; and

$$d_i(t) = d_0 + b \log\left(\frac{t}{i}\right) \tag{A5}$$

if $a = 0$ and $c = 0$. At time t, $1 - F_t(d)$ is then all of the nodes that have degree greater than d. If we solve for i such that $d_i(t) = d$, this then corresponds to the fraction of nodes that are older than i. That is, letting $i^*(d)$ be such that $d_{i^*(d)}(t) = d$, we then know that

$$1 - F_t(d) = \frac{i^*(d)}{t}.$$

¿From (A4) and (A5) we deduce that

$$i^*(d) = t\left(\frac{d_0 + \frac{b}{a} - \frac{ct}{1-a}}{d + \frac{b}{a} - \frac{ct}{1-a}}\right)^{\frac{1}{a}}. \tag{A6}$$

if $a > 0$ and either $c = 0$ or $a \neq 1$; and

$$i^*(d) = te^{\frac{d_0 - d}{b}} \tag{A7}$$

if $a = 0$ and $c = 0$. The claimed expressions for $1 - F_t(d)$ follow immediately. ▮

Lemma 3 *Consider an exponential growth process, where $n_t = (1+g)n_{t-1}$ and the degree of a node born at time i has initial degree d_0 and evolves according to*

$$\frac{dd_i(t)}{dt} = ad_i(t) + b.$$

Then the complementary cdf

$$1 - F_t(d) = \left(\frac{d_0 + \frac{b}{a}}{d + \frac{b}{a}}\right)^{\log(1+g)/a}. \tag{A8}$$

Proof of Lemma 3: The solution to

$$\frac{dd_i(t)}{dt} = ad_i(t) + b.$$

with initial condition $d_i(i) = d_0$ is

$$d_i(t) = \left(d_0 + \frac{b}{a}\right) e^{a(t-i)} - \frac{b}{a}.$$

This leads to a solution of

$$t - i^*(d) = \frac{1}{a} \log\left(\frac{d + \frac{b}{a}}{d_0 + \frac{b}{a}}\right),$$

where $i^*(d)$ is as defined in the previous proof. In the exponentially growing system with deterministic d_i, we have

$$1 - F(d) = \frac{n_{i^*(d)}}{n_t} = (1+g)^{-(t-i^*(d))}.$$

Substituting from the expression for $t - i^*(d)$ then leads to the claimed expression. ▌

Proof of Theorem 2:

Let us first derive the expression for $C^{TT}(g)$. Consider any give node i. Each i forms m new links. ¿From each node that i links to, there are m directed links. Thus there are m^2 possible pairs of directed links ij jk, and we need to determine the fraction of these where the link ik is present. To find the total number of such completed triples, we can alternatively simply count the number of situations where there is a pair of links ij and ik for which either jk or kj is present.

There are several situations to consider.

1. Both j and k were found at random.

2. One of j and k (say j) was found at random and the other by a network-based meeting.

3. Both j and k were found by network-based meetings.

In case 1, the probability of j and k being connected tends to 0 as n becomes large.

In case 2, we have $p_r(p_n m_n)$ such situations where the link to the random node was formed and then a link to a node in its neighborhood was formed; where $p_n m_n$ is the expected number of nodes formed through network-based meetings and on average p_r of them have a link from i to the parent node. [Other situations where k was not found through search of j's neighborhood, but instead through the search of some j''s neighborhood, will lead to a probability tending to 0 of the link jk being present.]

In case 3, if j and k were found by the search of different parents' neighborhoods, then the probability that they will be linked tends to 0. It is only in the case where they were found by search of the same parent's neighborhood that they will have a positive probability of being linked. There are on average $\frac{p_n m_n}{m_r} = \frac{p_r}{r}$ links formed by a new node to one of its parent's neighborhoods and m_r parents, and so there are $m_r \frac{p_n m_n}{m_r}(\frac{p_n m_n}{m_r} - 1)/2$ such pairs in total, in the situation where $\frac{p_r}{r} \geq 1$, and no such pairs otherwise (under the process described immediately before the theorem). As the parent and these links are independently and uniformly chosen, these potential clusters are completed with probability $\frac{C^{TT} m^2}{m(m-1)/2}$,[58] leading to approximately

$$\frac{C^{TT} m p_n m_n}{m-1} \left(\frac{p_r}{r} - 1 \right) \tag{A9}$$

completed triples from this case if $\frac{p_r}{r} \geq 1$, and 0 otherwise.

Thus, for a given node, summing across the three cases we expect

$$p_r p_n m_n + \frac{C^{TT} m p_n m_n}{m-1} \left(\frac{p_n m_n}{m_r} - 1 \right)$$

clusters out of m^2 possibilities if $\frac{p_r}{r} \geq 1$, and $p_r p_n m_n$ otherwise. Thus,

$$C^{TT} = \frac{p_r p_n m_n}{m^2} + C^{TT} \frac{p_n m_n}{m(m-1)} \left(\frac{p_n m_n}{m_r} - 1 \right), \tag{A10}$$

if $\frac{p_r}{r} \geq 1$, and

$$C^{TT} = \frac{p_r p_n m_n}{m^2}$$

otherwise. Solving for C^{TT} in (A10) yields the claimed expression.[59] We remark that $0 \leq C^{TT} \leq 1$ based on solving (A10). To see this, it can be checked directly that (A10) is of the form $C^{TT} = a + bC^{TT}$, where $0 < a \leq 1$ and $a + b \leq 1$.

Let us now derive the expression for C. Every completed triple is of the form ij, jk, ik, for some i, j and k. At time t there are t nodes and $m^2 C^{TT}$ such triples per node; for a total of $tm^2 C^{TT}$ triples. We only need find how this compares to the total number of possible pairs of relationships. Pairs come in three combinations (accounting for directions): ij ik, ij jk, and ji ki. There are $tm(m-1)/2$ of the first type, tm^2 of the second type, and $\sum_i d_i(t)(d_i(t)-1)/2$ of the third type. As each completed triple counts as a completion for three of the possible pairs, we can write

$$C = \frac{3m^2 C^{TT}}{m(m-1)/2 + m^2 + \frac{1}{t}\sum_i d_i(d_i - 1)/2}. \tag{A11}$$

¿From Theorem 1, the degree distribution of the process has a cumulative distribution function of

$$F_t(d) = 1 - \left(\frac{d_0 + rm}{d + rm}\right)^{\frac{m}{p_n m_n}},$$

and a corresponding density function of

$$f_t(d) = (rm)^{r+1}(r+1)(d+rm)^{-r-2}. \tag{A12}$$

Using the density function (and setting $d_0 = 0$), some straightforward calculations lead to an approximation of $\frac{1}{t}\sum_i d_i(d_i - 1)$ that is infinite if $r \leq 1$ and is $m(2mr + 1 - r)/(r-1)$ otherwise. Simplifying (A11) (noting that if $r > 1$ then it must be that $p_r/r < 1$ and so C^{TT} simplifies) then leads to the claimed expressions.

Finally, let us derive the expression for C^{Avg}. Again, using the density function from (A12) average clustering, $C^{Avg}(g)$, tends to

$$\int_0^\infty (rm)^{r+1}(r+1)(d+rm)^{-r-2}C(d)dd, \tag{A13}$$

where $C(d)$ is the clustering coefficient for a node with in-degree d.

We calculate $C(d)$ as follows. A node with in-degree d has

$$\frac{(d+m)(d+m-1)}{2} \tag{A14}$$

possible pairs of links that point in or out from d. This is the denominator of $C(d)$. The number of completed triples that involve the node is as follows.

First, there are situations where both links point out from the node i. As we discussed above, there will be approximately

$$C^{TT}m^2 \tag{A15}$$

such triples that are connected.

Next, there are situations where there is a link pointing in to i that was attached through the random process, and a link pointing out from i. First, we deduce that the number of such nodes that found i at random and have a link pointing into i (where i has degree d at time t) as follows. We know that this term $d_i^r(t)$ evolves according to

$$\frac{dd_i^r(t)}{dt} = \frac{p_r m_r}{t}$$

with initial condition $d_i^r(i) = 0$, and so (A5) tells us that

$$d_i^r(t) = p_r m_r \log\left(\frac{t}{i}\right)$$

Then from equation (A6), from the process of $d_i(t)$ (not to be confused with $d_i^r(t)$), it follows that

$$\frac{t}{i} = \left(\frac{d + \frac{1}{rm}}{\frac{1}{rm}}\right)^{\frac{m}{p_n m_n}}.$$

Combining these two equations, we deduce that

$$d_i^r(d) = rm\left[\log\left(\frac{d}{rm} + 1\right)\right],$$

where $d_i^r(d)$ is the number of inward links that were formed through the random process to a node with in-degree d. Each of the nodes that found i through random meetings has $p_n m_n / m_r$ links that are attached based on a search of i's neighborhood, and thus we have

$$rm\left[\log\left(\frac{d}{rm} + 1\right)\right]\frac{p_n m_n}{m_r} \tag{A16}$$

such completed triples.

The remaining types of pairs of links that involve i are: ones where there is a link pointing into i that was formed by a network-based meeting, with another link pointing out from i, and ones where there are two links pointing into i at least one of which was formed through a network-based meeting. In any such situation where there is a completed triple, one of the nodes, say j, has links ji and jk, where ji is formed through a network-based meeting. We can simply add the expected number of this type of completed triple over each node j that has attached to i through a network-based meeting. First, there is a p_r chance that j will have attached to the parent through whom j located i (and necessarily the third link is then present). The other potential triples that will occur with a non-trivial probability in the limit are those where j has connected to i through a network-based meeting and also to some other node k through a network-based meeting. A fraction $\frac{2}{p_n m_n}$ of the potential pairs of outward links from j that are both formed through network-based meetings will involve a link to i. From the cases 1, 2, and 3, above, we know that these fit into the third case and can be calculated by looking at the $C^{TT} m^2$ and subtracting off the numbers from the first two cases, leading to a total of

$$\frac{2}{p_n m_n}\left(C^{TT} m^2 - p_r p_n m_n\right)$$

such triples. Given that there are $d - d_i^r(d)$ nodes that found i through network-based meetings, and substituting for $d_i^r(d)$, we obtain the expression

$$\left(d - rm\left[\log\left(\frac{d}{rm}+1\right)\right]\right)\left(p_r + \frac{2}{p_n m_n}\left(C^{TT}m^2 - p_r p_n m_n\right)\right) \tag{A17}$$

for the number of completed triples of this type.

Finally, by summing (A15), (A16), and (A17), we find the numerator of $C(d)$, and (A14) provides the denominator. Plugging this expression for $C(d)$ into (A13) provides the claimed expression for C^{Avg}. ▮

Proof of Theorem 4: If $d_i(t) > d_j(t)$, then under the mean-field approximation, if we let i and j be the birth dates of those nodes, then it must be that $i < j \leq t$. Next note that for $d < d_i(t)$,

$$1 - F_i^t(d) = \frac{d_i(i_t^*(d))}{d_i(t)},$$

where $i_t^*(d)$ is the date of birth of a node that has degree d at time t; and for $d \geq d_i(t)$

$$1 - F_i^t(d) = \frac{0}{d_i(t)}.$$

Thus, we need only consider $d < d_j(t)$, as the result is clear for $d \in [d_j(t), d_i(t))$. It is thus enough to show that for any $i < j < t' < t$

$$\frac{d_i(t')}{d_i(t)} > \frac{d_j(t')}{d_j(t)}.$$

This is easily verified by direct calculations from (A4). ▮

Proof of Theorem 5: From the proof of Theorem 2, we have

$$C(d) = \frac{m^2 C^{TT}\left(1+\frac{2d}{p_n m_n}\right) - p_r d + rm\left[log\left(\frac{d}{rm}+1\right)\right]\left(\frac{p_r}{r}+p_r-\frac{2C^{TT}m^2}{p_n m_n}\right)}{(d+m)(d+m-1)/2}$$

Thus $C(d)$ is approximated by $\frac{\left(\frac{2m^2C^{TT}}{p_n m_n}-p_r\right)d}{\frac{1}{2}d^2}$ for large d. If we show the expression $\frac{2m^2C^{TT}}{p_n m_n}-p_r$ to be positive, then the approximation is decreasing in d, which implies the result. Let us check that $\frac{2m^2C^{TT}}{p_n m_n}-p_r$ is positive. We know from Theorem 2 that $C^{TT} \geq \frac{p_r}{m(1+r)}$. Thus,

$$\frac{2m^2C^{TT}}{p_n m_n} - p_r \geq \frac{2mp_r}{(1+r)p_n m_n} - p_r = \frac{2mp_r}{m} - p_r > 0,$$

and so the expression is positive. ▮

Proof of Theorem 6: By standard results on second order stochastic dominance (e.g., Michael Rothschild and Joseph Stiglitz (1970)), it is sufficient to show that

$$\int_{d_0}^{X} [F(d) - F'(d)] dd > 0 \tag{A18}$$

for all $X > 0$. Substituting from (4), we rewrite (A18) as

$$\int_{d_0}^{X} \left[\left(\frac{d + r'm}{d_0 + r'm} \right)^{-1-r'} - \left(\frac{d + rm}{d_0 + rm} \right)^{-1-r} \right] dd.$$

Setting $d_0 = 0$ and integrating, we obtain

$$-m \left(\left[\left(\frac{X + r'm}{d_0 + r'm} \right)^{-r'} - 1 \right] - \left(\frac{d_0}{rm} + 1 \right) \left[\left(\frac{X + rm}{d_0 + rm} \right)^{-r} - 1 \right] \right).$$

It is sufficient to show that

$$\left(\frac{X}{r'm} + 1 \right)^{r'} > \left(\frac{X}{rm} + 1 \right)^{r},$$

or that $\left(\frac{X}{rm} + 1 \right)^{r}$ is increasing in r. It is thus sufficient to show that the log of the same expression is increasing in r. Taking the log and then differentiating leads to a derivative of

$$log \left(\frac{X}{rm} + 1 \right) - \frac{\frac{X}{rm}}{\frac{X}{rm} + 1}.$$

This expression is 0 when $X = 0$, and is strictly increasing in X (the derivative of this expression with respect to X is clearly positive at $X > 0$), and so is positive whenever $X > 0$. ▮

Notes

[0] Jackson is at the Department of Economics, Stanford University, Stanford, California 94305-6072, USA, email: jacksonm@stanford.edu and webpage: http://www.stanford.edu/∼jacksonm . Rogers is at MEDS, Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60208, email: b-rogers@northwestern.edu and http://www.kellogg.northwestern.edu/faculty/rogers_b/personal/. We gratefully acknowledge financial support under NSF grant SES-0316493, the Lee Center for Advanced Networking, and a SISL/IST fellowship. We thank David Alderson, Hawoong Jeong, Sanjeev Goyal, Marco van der Leij, and José Luis Moraga-González for making data available. We thank Steven Durlauf for a helpful discussion

of the paper and Antoni Calvo-Armengol, Matthias Dahm, Dunia Lopez-Pintado, Fernando Vega-Redondo, and Duncan Watts for helpful comments and conversations, as well as Douglas Bernheim, Vince Crawford, and three anonymous referees for helpful suggestions.

[1] The term "social" emphasizes that links result from the decisions of nodes or agents themselves, and are not designed or optimized, as in the case of a technological network or biological system.

[2] We add a caveat that claims of such regularities appearing in the literature are based on an accumulation of case studies. There is no work that systematically looks across networks to carefully document the extent of these facts. As discussed below, one contribution of the model is that it opens the possibility for such a study.

[3] This stylized fact is captured in the famous "six degrees of separation" of John Gaure's play. Stanley Milgram (1967) pioneered the study of path length through a clever experiment where people had to send a letter to another person who was not directly known to them. The diameters of a variety of networks have been measured varying from purely social networks, to co-authorship networks, to parts of the internet and world wide web. See Albert-László Barabási (2002) for an illuminating account.

[4] Ideas behind clustering have been important in sociology since Georg Simmel (1908) who pointed out the interest in triads. An important recent account of clustering is Duncan Watts (1999).

[5] These distributions date to Vilfredo Pareto (1896) and have appeared in a variety of settings ranging from income and wealth distributions, distribution of city populations, to degree distributions in networks (e.g., Derek J. deSolla Price (1965)). For an informative overview, see Michael Mitzenmacher (2004).

[6] While it appears that some observed empirical degree distributions are closer to "scale-free" than random, there is remarkably little careful statistical testing to establish which distributions actually fit the data. David M. Pennock et al (2002) is the only work other than ours to provide fits across applications (to the best of our knowledge). "Eyeballing" the data is a particularly inappropriate (although regularly used) technique, since distributions such as the Pareto and lognormal distributions are nearly indistinguishable visually for many parameters on a log-log plot. As we shall see when we fit different data sets, degree distributions are remarkably varied.

[7] Positive assortativity appears to be special to social networks. Negative correlation is more prevalent in technological and biological networks (e.g., see Mark E. J. Newman (2003)). The term "assortative" has been used in the networks literature, while the shorter term "assortive" is occasionally used in the matching literature.

[8] See Sanjeev Goyal, Marco J. van der Leij, and José Luis Moraga-González (2003). This can also be seen in the data reported in Table II in Newman (2003). He reports two different clustering measures for several networks. One is an average of local clustering across nodes, and the other is an overall clustering. The latter statistic is smaller in each case. As the average clustering under-weights high degree nodes, this is suggestive of such a negative relationship.

[9] We show a strong version of this in the form of stochastic dominance.

[10] See Matthew Jackson (2004) for a recent survey.

[11] See Newman (2003) for a survey. See Jackson (2007) for a discussion comparing the economics approach and the random graph approach.

[12] The strategic network formation literature also contains explanations of the small world phenomenon. See Nicolas Carayol and Pascale Roux (2003), Andrea Galeotti, Goyal and Jurjen Kamphorst (2004), and Jackson and Brian Rogers (2005).

[13] The logic behind this traces back to early explanations of power laws due to G. Udny Yule (1925) and Herbert Simon (1955). Simon argued that in a growing population, if individual object size (here, degree) grows according to a lognormal distribution over time, and subject to some bound on object size (here, zero), then the overall distribution of object size in the population will have a scale-free distribution. See Harry Kesten (1973) for a formal treatment and Mitzenmacher (2004) for an overview.

[14] "HOT" (highly optimized tolerance) systems are centrally optimized rather than self-organizing. As such, the explanation for fat-tailed degree distributions is different both in application (for instance, understanding connections among some routers) and approach, and thus complementary to the model proposed here. Also, HOT systems deliberately do not exhibit some of the other features discussed here.

[15] As a separate point, many previous models involve artificial rewiring or behavior that might be hard to rationalize. The network-based meeting model that we present is a natural behavior that not only is easy to envision but is actually part of many (approximately) optimal algorithms.

[16] For other variations and extensions of preferential attachment see Sergey N. Dorogovtsev and José F. Mendes (2001), Mark Levene et al (2002), and Cooper and Frieze (2003)).

[17] The out-degree of a node i is simply the number of links from i to other nodes. We concentrate on in-degree. In a non-directed network, in-degree and out-degree coincide and are simply referred to as the node's degree.

[18] In a case where the union of the parents' neighborhoods (excluding the parents) consists of fewer than m_n nodes, then all of those and only those nodes are picked.

[19] For example, the models of Price (1976) and Barabási and Albert (1999), can be captured by setting $m_r = m_n = p_n = 1$ and $p_r = 0$; and variants on Paul Erdös and Alfréd Renyi's (1960) random graphs (e.g., Callaway et al (2001)) are cases where $p_n = m_n = 0$.

[20] (1) is not an exact calculation, since it ignores the possibility that some of the parents are in each others' neighborhoods, or that the union of the parents' neighborhoods has fewer than m_n total members. Nevertheless, it is a good approximation when the network is large (i.e., t is large) relative to m_r and m_n, and when m_n is small compared to $m_r(p_r m_r + p_n m_n)$, as these adjustments vanish. When m_n becomes large relative to $m_r(p_r m_r + p_n m_n)$, then the probability that i is found out of the parents' neighborhoods once a neighbor of i is found at random goes to 1, and so $\frac{m_n}{m_r(p_r m_r + p_n m_n)}$ is bounded above by 1 and the calculations are an easy extension of those presented here.

[21] For the case of pure preferential attachment it is necessary to start the in-degree at a level different from

0, or a node would never get any links.

[22] This presumes that $p_n m_n > 0$, as otherwise (3) simplifies and has a different solution, as discussed in the appendix.

[23] While the realization in the first panel of Figure 1 does not offer a precise match, in that case the model simplifies and one can verify the degree distribution exactly (e.g., see Béla Bollobás et al (2002)); the difference is due to the noise in the simulations.

[24] Each panel depicts the results of a single computer simulation.

[25] To have pure preferential attachment rather than network-based meetings, (3) should be rewritten as $\frac{dd_i(t)}{dt} = \frac{md_i(t)}{tm+td_0}$, since d_0 matters in the degree count. The corresponding solution for the complementary cdf (see the appendix for details) is then $1 - F(d) = d^{-\frac{m+d_0}{m}}$. Setting $d_0 = m$, this corresponds to the the $\Pr(d) \sim d^{-3}$ of Barabasi and Albert (1999).

[26] When $r \to \infty$, (5) is no longer well-defined. At that extreme (see the appendix for details), $\log(1 - F(d)) = \frac{d_0 - d}{m}$, which is an exponential distribution. Since this is a growing system the degree distribution differs from the static random graphs analyzed by Erdös and Rényi (1960).

[27] "Clustering" has several different connotations in the sociology literature. We follow the terminology from the recent literature on large networks (e.g., see Newman (2003)).

[28] In the directed version of our model, it is not possible to have a directed cycle, as nodes only form directed links to nodes born at earlier dates, and so we do not need to worry about k having a directed link to i.

[29] See Table II in Newman (2003) for some illustration of the differences between $C(g)$ and $C^{Avg}(g)$.

[30] To see that the average clustering coefficient is positive, note that a lower bound on $m^2 C^{TT}$ is $p_r p_n m_n$. Then a lower bound on the integral is $\int_0^\infty f(d) \left(\frac{p_r p_n m_n + d p_r}{(d+m)(d+m-1)} \right) dd$, where $f(d)$ is the density function $(rm)^{r+1} (r+1) (d + rm)^{-r-2}$. This can be directly verified to be positive when $r > 0$ and $m \geq 1$.

[31] This does not show up in the fraction of transitive triples calculations since these sorts of link pairs (both pointing in to a given node) are not involved in that calculation.

[32] Given the directed nature of the links, there are various ways to measure diameter. We let the distance between two nodes be the minimum number of links needed to pass from one node to the other when ignoring the direction of the links. Clearly, the diameter will be infinite if we require that each node reach every other by a directed path, because directed paths always point in the direction of increasing age.

[33] We need to allow nodes to self-connect and enter as if they had degree 1, in order to directly apply their proof. Self-connections can be added and then ignored in interpreting the network.

[34] For example, we see assortativity in models by Krapivsky and Redner (2002) as well as Callaway et al (2001).

[35] Note that through the in-degree relationships, one can infer corresponding out-degree relationships.

[36] If $d \geq d_i(t)$, then it is clear that $1 - F_i^t(d) = 1 - F_j^t(d) = 0$, as then d corresponds to nodes that are older than both i and j.

[37] This is complicated by the fact that if i is found via a random meeting, then its relatively larger neighborhood size means that its neighborhood tends to receive more of the newborn node's links. However, as a node's degree grows, the relative fraction of new links it gets comes increasingly from network-based meetings.

[38] We obtained the www data from Albert, Jeong, and Barabási (1999), the co-authorship data from by Goyal, van der Leij, and Moraga-González (2003), the citation network from Eugene Garfield, the prison data from Duncan MacRae (1960), the ham radio data from Peter D. Killworth and H. Russell Bernard (1976), and the high school romance data from Peter Bearman, James Moody, and Katherine Sovel (2004).

[39] Hypothetically, one could estimate m and r using Maximum Likelihood techniques. However, that would entail being able to derive a formula for the likelihood of observed data as a function of the parameters. While we have a closed form for the degree distribution (under a mean-field approximation), deriving the probability of seeing arbitrary degree distributions on any number of nodes is generally intractable. Even using simulation techniques, deriving relative likelihoods for different degree distributions directly is intractable since it involves exponentially many calculations. There may be useful approximation algorithms, but we leave that to further research as the approach followed here is computationally easy and generates excellent fits.

[40] This iteration converges to the same fixed point whether we start from very high or very low guesses of the initial $\widehat{r}$.

[41] The only reason for allowing for differences in p_r and p_n in the model was to be able to nest other models from the previous literature such as the pure preferential attachment model, which requires $p_n \neq p_r$.

[42] One could alternatively estimate p from examining the variance in out-degree. As this is not available from all the data sets, and since clustering is more central to our analysis, we have chosen to fit it from the clustering calculations.

[43] The prison friendship and high school romance data could also be interpreted to be non-directed. However, those latter two data sets are based on surveys and one person can name another, without the second naming the first. As such, the data are in the form of a directed graph and so we treat these directly from the model. This presumes that the person naming the other as a friend is the person who initiated the friendship. As this might not be the case, we could also fit these data as non-directed networks. This makes little difference in the estimates.

[44] Some papers in the co-authorship data set involve three or more researchers, which is a type of link that we do not model. As only 11 percent of the papers involve three or more authors, we ignore this complication in our fitting of our model to that data set. This would be more problematic for fitting collaboration networks in some other disciplines, where the typical number of co-authors on papers is much larger than two.

[45] The fits for the average clustering for the Prison and Ham networks are done based on simulations. In estimating the average clustering we have entered 0's when a node has a degree of 0 or 1. If we simply omit these nodes, then the averages increase (to .33 for the prison data, and .69 for the Ham data). We use

a similar convention for 0-degree nodes when estimating degree correlations. The asymptotic estimates of $C^{avg}(g)$ based on Theorem 2 are $.001p$ for the prison network and $.09p$ for the Ham network. Since these estimates are developed based on asymptotically large networks we resort to simulations, and hence the fits are less precise than for the larger networks. For the other networks, the number of nodes is large enough for the asymptotic estimates to be accurate.

[46]A proper modeling of this network would include different sexes and have most nodes link to nodes of the opposite sex. While this might change the specifics of the degree distribution, it would have similar features in that it would range between something that was uniformly random and something that looked more like preferential attachment. The fact that the fit of our degree distribution matches a random network suggests that the fit would be the same in a more detailed model.

[47] Newman (2003) reports a figure of .29, referencing Albert, Jeong, and Barabási (1999) and Barabási, Albert, Jeong (2000), although we cannot find such a figure in those articles and have not been able to obtain the full data set to estimate it. To match that figure would require p closer to 1.

[48]Given an r of ∞, the number of attachments formed through network-based meetings is negligible, and since we need to set m_r and m_n to be integers for the simulations, we approximated the estimates with $m_n = 0$ for these simulations. For the Ham data set, as the ratio is $r = 5.0$ and $m = 3.5$, and we need to set m_r and m_n to be integers, we used $m_r = 5$ and $m_n = 1$, and then set $p = m/(m_r + m_n)$. Thus, for the Ham simulations we must use a p that is lower than the estimated p.

[49] If one instead fits the data presuming that it is log-log linear, rather than through our model, then one obtains a pdf of $f(d) \sim d^{-2.56}$ (the corresponding coefficient on the cdf is then -1.56). This differs slightly from the coefficient of -2.1 reported in Albert, Jeong and Barabási (1999), where the data is coalesced into bins before fitting the model.

[50]For high enough values of d_0, the result does not hold. As $d_0 > 0$ is not really an interesting case, but simply included to allow us to nest pure preferential attachment as a well-defined special case, the case of $d_0 = 0$ is the relevant one.

[51]This result has an obvious variation for the case where expected utility is strictly convex in degree, in which case the ordering is simply reversed.

[52]An example of a ‘connections’ model where utility is derived from indirect connections was studied by Jackson and Asher Wolinsky (1996). (See Jackson (2004) for a survey of the related literature.)

[53]This was not critical in the directed case, as it was only the *out*-degree of the parent nodes that was important in the network-based meeting process and this was i.i.d. across nodes.

[54]This is often known as George K. Zipf’s law (1949), even though Zipf was concerned with many things including word usage. See Xavier Gabaix (1999) for a recent model of city growth and a discussion of Zipf’s law.

[55]We thank David Alderson for sharing these data with us.

[56]We remark that the literature often focuses on the largest areas in terms of populations (for instance,

Gabaix's (1999) Figure 1 only includes the 135 largest cities), and hence would be looking mostly at the tail of the distribution, which would be consistent with Zipf's law.

[57] Also, note that in the log-log plot, *90 percent* of the counties have log size less than 8.7. Thus, the majority of the graph itself (the part which is 'most linear') is generated by only 10 percent of the data. In fact, Murray Gell-Mann (1994) suggests that a function similar to the degree distribution that we have derived would would better match the data.

[58] A given parent i' has approximately $C^{TT}m^2$ completed triples of $m(m-1)/2$ possible pairs of outward links.

[59] Note that we have done this calculation for a typical i, and so this confirms our earlier claim that the overall and per node average version of the fraction of transitive triples coincide.

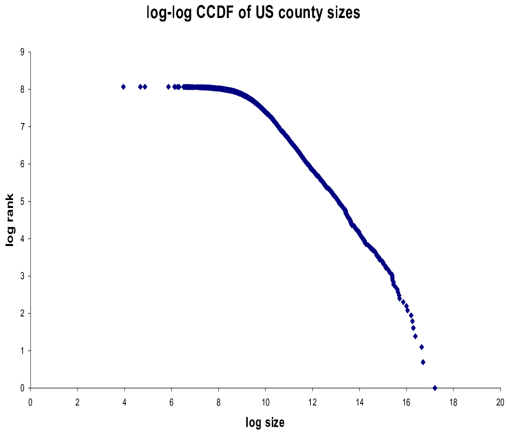

Figure 3. *log-log plot of the complimentary cdf of US county sizes showing that while the tail of the distribution is approximately scale free, the remainder of the data (over 90%) is not.*