Search in the Formation of Large Networks: How Random are Socially Generated Networks?

by Matthew O. Jackson and Brian W. Rogers *

May 2004
revised: January 4, 2005

Abstract

We present a model of network formation where entering nodes find other nodes to link to both completely at random and through search of the neighborhoods of these randomly met nodes. We show that this model exhibits the full spectrum of features that have been found to characterize large socially generated networks. Moreover, we derive the distribution of degree (number of links) across nodes, and show that while the upper tail of the distribution is approximately "scale-free," the lower tail may exhibit substantial curvature, just as in observed networks. We then fit the model to data from six networks. Besides offering a close fit of these diverse networks, the model allows us to impute the relative importance of search versus random attachment in link formation. We find that the fitted ratio of random meetings to search-based meetings varies dramatically across these applications. Finally, we show that as this random/search ratio varies, the resulting degree distributions can be completely ordered in the sense of second order stochastic dominance. This allows us to infer how the relative randomness in 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.

*The Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, USA, emails: jacksonm@hss.caltech.edu and rogers@hss.caltech.edu, web sites: http://www.hss.caltech.edu/∼jacksonm/Jackson.html and http://www.hss.caltech.edu/∼rogers. 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.

1 Introduction

Network structures play a central role in determining outcomes in many settings. Examples include the world wide web, co-author relationships among academics, joint research venture projects among firms, political alliances, trade networks, the organization of intra-firm management, the sharing of job opening (and other sorts of) information through social networks, and P2P systems for file sharing, among others. Given the large and increasing prevalence of situations where network structure plays a key role, it is important to understand the process that determines network form, as well as to understand the efficiency properties of emerging networks and how that relates to the formation process.

In this paper we provide a model of network formation that does several things. First, it leads to networks that exhibit characteristics that have been found to be 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 model to data and back out underlying parameters of the model. Third, we show how the efficiency of the emerging networks vary as the parameters of the model are varied. Before describing the model, let us provide some brief background on what is known regarding large socially-generated networks.

1.1 Characteristics of Socially-Generated Networks

Key empirical regularities that have been found to be shared by socially-generated networks can be summarized as follows.[1]

(i) Such networks tend to have small diameter and small average path length, where small is on the order of the log of the number of nodes or less.[2]

(ii) Such networks tend to have high clustering coefficients relative to what would emerge

[1] We add a caveat that the claims of such regularities that appear 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 we discuss below, part of the contribution of our model is that it opens the possibility for such a study.

[2] The diameter is the maximum distance between any two nodes of the network, where distance between nodes is defined as the shortest path between them measured in the number of links or edges. This stylized fact is captured in the famous "six degrees of separation" of John Gaure's play. Stanley Milgram [45] 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 Barabási [6] for an illuminating account.

if the links were simply determined by an independent random process.[3]

(iii) The degree distributions of such networks tend to exhibit "fat tails" and in some cases approximate a "scale-free" or "power-law" distribution, at least in the upper tail.[4] Thus, there tend to be many more nodes with very small and very large degrees than one would see if the links were formed completely independently.

(iv) The degree of linked nodes tends to be positively correlated in socially generated networks. This is referred to as assortivity and appears to be special to networks where the links are formed by the decisions of people controlling individual nodes, and contrasts with the opposite relationship that is more prevalent in technological and biological networks (e.g., see Newman [47]).

(v) In such networks, the clustering among the neighbors of a given node 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 as compared to the neighbors of a lower degree node.[5]

While these characteristics are far from enough to completely characterize a network, together they give us a great deal of information about network structure.

1.2 A Preview of Our Model and Results

Let us 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 completely at random. Second,

[3]Clustering coefficients look at two neighbors of a given node and ask what is the frequency with which they tend to be connected to each other. Heuristically, what is the chance that my friends are friends of each other? Ideas behind clustering have been important in sociology since Simmel [54] who pointed out the interest in triads. An important recent account of clustering is Watts [58].

[4]The degree of a node is simply the number of links that involve that node as one of the endpoints. One can distinguish between inward links and outward links in the case of a directed network. An example of a power law distribution is the Pareto distribution, where the relative frequency of nodes with a degree of d is proportional to $d^{-\gamma}$ for some $\gamma > 1$. These distributions date to Pareto [49], and have appeared in a wide variety of settings ranging from income distributions, distribution of city populations, to degree distributions in networks. For an informative overview, see Mitzenmacher [46].

[5]See Goyal, van der Leij, and Moraga-González [28]. This can also be seen in the data reported in Table II in Newman [47]. 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 shows that there must be some sort of negative relationship.

it then meets some of those nodes' neighbors. This second process is what we refer to as "search." 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 heuristically why this process exhibits features (i)-(v). Nodes with higher degree are more likely to be found through the local search process since more paths lead to them. This leads to attachment that has characteristics similar to preferential attachment, which in turn leads to "scale-free"-like characteristics for the upper tail of the degree distribution. However, the combination of this with random meeting also allows for richer distributions that can exhibit non scale-free characteristics more generally. As this is a growing network, it will exhibit assortativity since older nodes are more likely to be linked to each other, and more likely to have large degree. We show a strong version of this in the form of stochastic dominance. To understand why this process leads to high clustering, note that the local search process leads to a tendency to form local links. For instance, a link might be formed to the point of entry and then also to a node found by searching along a link from that point of entry. This naturally leads to high clustering that does not disappear as the network becomes large. The negative relationship between node degree and local clustering comes partly from the fact that the potential number of triads goes up with the square of degree while the probability of forming triads goes up approximately linearly in degree. The relatively small diameter comes from the tendency for many nodes to find the same ones to link to (as nodes are more likely to find and link to nodes which have large numbers of links), and also for these nodes to at the same time link randomly to other neighborhoods, which in turn generates a diameter which is of smaller order than that of a either purely random network or one with single-link preferential attachment.

A key feature of the model is that as the relative roles of search and random attachment are varied, the specifics of network characteristics, such as degree distribution and clustering coefficients, change in ways that we characterize. This allows us to fit the model to data and infer the relative rates of random attachment and search in various applications. In fitting the data to six different networks, we find widely different ratios of the role of random attachment to search. Although the model is very simple, it fits the data remarkably closely 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 search process compared to random meetings is roughly seven times greater in a world wide web application than it is in a co-authorship network, and is almost completely random in two friendship applications.

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, we show that as the ratio of random attachment to search is varied, we can completely order the degree distribution in the sense of second order stochastic dominance. This turns out to be very useful as it allows us to derive results about the efficiency of the resulting networks. For instance, if the utility provided 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.

1.3 Relation to the Literature

There is extensive study of networks in a number of disciplines, including sociology, economics, computer science, and statistical physics. The sociology literature provides a very rich background on social networks, with numerous case-studies including most of the early studies underlying the stylized facts listed above. Formal modeling of network formation lies mainly in the economics, computer science, random graph, and statistical physics literatures. Such models 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.[6] The other set is more mechanical, describing stochastic processes of network formation that are meant to exhibit some set of characteristics. This set has roots back in the early random graph literature, and has been very recently flourishing in the computer science and physics literatures.[7]

Our model has characteristics of both of these categories. On the one hand, the nodes or agents in our model are non-strategic. While this meeting process has many "natural" characteristics, it is more in the tradition of the random graph literature in terms of being a model 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 nodes, and thus deduce some efficiency characteristics of the networks. Such efficiency issues have been pretty much exclusively the realm of the economics literature.

A variety of recent random graph models have been proposed to explain some of properties (i) to (v).[8] For example, Watts and Strogatz ([59], [58]) generate networks exhibiting small-

[6] See Jackson [30] for a recent survey.

[7] See Newman [47] for a survey.

[8] There is also a small part of the strategic network formation literature that explains "small worlds" phenomena. See Carayol and Roux [13], Galeotti, Goyal and Kamphorst [25], and Jackson and Rogers [31].

world characteristics,(i) and (ii), by starting with a highly regular and symmetric network and randomly rewiring some links. Barabási and Albert and others ([7], [16]) have shown that networks with scale-free 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).[9] Scale-free distributions have also been shown to result if new nodes copy the links of a randomly identified node (Kleinberg et al [35] and Kumar et al [39]),[10] or if networks are designed to optimize tolerance (e.g., Carlson and Doyle [14] and Fabrikant et al [20]).[11] A variation on preferential attachment where only some nodes are active at any time (Klemm and Eguíluz [36], [37]) has been shown to also exhibit small-world properties (i)-(ii). And, some network models that grow over time have been shown to exhibit (iv) (e.g., Callaway et al [12] and Krapivsky and Redner [38]).

While the above described models made important progress in helping us to understand some of the specific empirical regularities of large networks, none of these previous models are 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.[12]

Our simple model exhibits all of the stylized facts by combining random meetings and network search in a natural way. While the search aspect of our model is easily seen to generate a form of preferential attachment, it is important to understand that the particular relationship between random meetings and search here is critical to obtaining our results. Models that mix random meetings and preferential attachment (e.g., vertex copying – see Kleinberg et al [35] and Kumar et al [39]) or have nodes randomly decide to form their links

[9]The logic behind this traces back to early explanations of power laws due to Yule [60] and Simon [55]. Simon argued that in growing population, if individual object size (say degree) grows according to a lognormal distribution over time, and subject to some bound on object size, then the overall distribution of object size in the population will have a scale-free distribution. Preferential attachment produces relative growth that is proportional to size, as occurs with lognormal growth. See Kesten [33] for a formal treatment and Mitzenmacher [46] for both the history and an overview of some of the arguments.

[10]There are many other studies generating scale free degree distributions based on variations of preferential attachment, including some models that are hybrids of random and preferential attachment (e.g., see Dorogovtsev and Mendes [17], Levene [17], Levene et al [40], Pennock et al [51], and Cooper and Frieze [16]).

[11]That important idea of "HOT" (highly optimized tolerance) systems examines the implications of systems are centrally optimized, rather than self-organizing. As such the explanation is quite different both in application (for instance, understanding connections among some routers) and approach, and thus complementary to the model proposed here; also such designed HOT systems will generally (deliberately) not exhibit some of the other features discussed here.

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

one way or the other (see Pennock et al [51]), cannot exhibit all of the features discussed here. In particular, we obtain the high clustering (ii), a diameter that is smaller than 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 to neighborhoods resulting from that random search.

The most similar models in structure to ours are by Vazquez [56] and Pennock et al [51]. Vazquez's model has nodes enter by finding a random node, and then sequentially choosing randomly between either following a path from the last node visited, or randomly jumping to a new node. While the model differs from ours in details, it is similar in that it combines random meetings with some local search of neighborhoods. Although the models are similar in these regards, the only overlap in our analyses is in showing a nonzero clustering. The model of Pennock et al [51] adds random link formation to a preferential attachment model. This leads to a degree distribution that is similar to ours. However, there is no limiting clustering in such a model, and other characteristics of their model are also quite different.

In order to clarify one of the differences of our work from previous work, it is necessary to expand upon what is known about degree distributions, (iii). While it is clear that the degree distribution of most observed social networks differ significantly from a purely random network, it is not clear exactly what the degree distributions really are (see Mitzenmacher [46]). As pointed out by Pennock et al [51], many of the internet based data sets that are said to be "scale-free," are only scale-free for large degree nodes.[13] While it appears that many observed empirical degree distributions are closer to "scale-free" than random, there is remarkably little careful statistical testing to establish what distributions actually best fit the data. "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. (Recall that on a log-log plot a small fraction of the data points end up occupying most of the area of the graph.)

An important advantage of our work is that it results in a family of degree distributions that span from negative exponential, as occurs in the purely random case, to purely scale-free. As we fit the model to data, we provide the first actual fits of degree distributions to observed networks that we are aware of. Interestingly, we find that the degree distributions vary substantially across applications and, moreover, that networks that have been "scale-free" in the literature, in fact differ significantly (in a well-defined way) from being scale-free.

Finally, as mentioned above, results relating variations in the formation process to variations in efficiency do not appear in the previous literature.

[13] Alderson [4] gives several arguments for why the internet may be less scale-free than has been claimed.

2 The Model

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 the case where $g_{ij} = g_{ji}$ for all nodes i and j.

For any node $i \in N$, let $d_i(g) = |\{j \in N \,|\, g_{ji} = 1\}|$ denote the in-degree of i. In a non-directed network, degree and in-degree will coincide. Also let $n_i(g) = \{j \in N \,|\, j_{ij} = 1\}$ denote i's neighborhood.

The model is based on a process through which nodes meet each other. Action takes place at a countable set of dates $t \in \{1, 2, \ldots\}$. At each time t a new node is added to the population. Let N_t denote the set of all nodes present at time t. Denote by $g(t)$ the network consisting of the links formed on the nodes N_t at the end of time t.

The formation of links is described as follows. Let us denote the new node born at time t by t. Upon birth, the node t identifies m_r nodes uniformly at random (without replacement) from N_{t-1}. We shall call these "parent" nodes. The new node forms a (directed) link to a given parent node if the benefit in terms of utility from forming that link, exceeds the cost. For now, let us assume that the benefit less the cost is independently and identically distributed across t, t' pairs, regardless of the network structure. Let p_r denote the probability that a new node finds a randomly identified node attractive to link to. In section 5.1 we return to richer formulations of utility that allow for indirect benefits and externalities from the network structure.

In addition, (regardless of whether the node forms a link to the parent) the node t searches the parents' neighborhoods and finds other nodes. For instance, in the example of web pages, new nodes are found by following links from the parents' web pages. The new node t finds m_s nodes through this search method (over all parents). We think of this as happening in the parents' immediate neighborhood, but the same analysis applies for more extended neighborhoods - for instance, searching along paths of length at most k from the parent node. Let p_s denote the probability that the new node obtains a positive utility from linking to a given node found through search.[14]

Generally, it is reasonable to have $p_r = p_s = p$, but we allow for the additional heterogeneity so that we can nest other models as special cases.[15]

[14] In order to have the process well-defined upon starting, simply start with $(m_s + m_r)^2$ nodes, where each of them each have $(m_s + m_r)$ neighbors (who are otherwise unconnected).

[15] For example, Barabási and Albert [7], who model the case of pure preferential attachment, have $m_r =$

An expression for the probability that a given existing node i with degree[16] $d_i(t)$ gets a new attachment (in period $t+1$) is roughly[17]

$$p_r \frac{m_r}{t} + p_s \left(\frac{m_r d_i(t)}{t} \right) \left(\frac{m_s}{m_r(p_r m_r + p_s m_s)} \right), \tag{1}$$

The first expression in (1) is the probability that the node is chosen at random as a parent of the new node and is linked to in that capacity. There are t existing nodes, and a new node picks m_r of them at random. The second probability is that the node is found and attached to via the search. This is the probability that at least one of the nodes that has a link to i is chosen as a parent, times the probability that i is then found via the search, and then attached to. There are $d_i(t)$ possible nodes that would have i in their neighborhood, and so the probability that one of them is identified as a parent by the new node is $\frac{m_r d_i(t)}{t}$, and then the corresponding probability that the node is identified out of the search through the neighborhoods of the parents is $\frac{m_s}{m_r(p_r m_r + p_s m_s)}$.

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

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

2.1 A Mean-field Analysis of the Degree Distribution

The analysis of this random system is complicated, given the combination of both random attachment and a search that depends on the structure of the previous graph. Thus, we use combinations of techniques that are common to analyzing such dynamic systems. First, we analyze a "mean-field" approximation to this system. This is a continuous time system where all decisions happen for certain at a rate proportional to the expected change. Second, we run some simulations of the actual system and compare these to the predictions of the mean-field approximation. Third, we compare the results to closed form solutions for some special extreme cases.

$m_s = p_s = 1$, and $p_r = 0$; and variants on Erdös and Renyi [19] (e.g., Callaway, Hopcroft, Kleinberg, Newman and Strogatz [12]) are cases where $p_s = m_s = 0$.

[16] In the context of our model, we use this term to mean in- degree, since out-degree is homogenous across nodes.

[17] This is not an exact calculation, since it ignores the possibility, for instance, that some of the parents are in each others' neighborhoods, or that a node is found by more than one method of search. Nevertheless, it is a very accurate approximation when the network is large (i.e., t is large) relative to m_r and m_s, as these adjustments vanish.

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

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

If we start the system with each node t having in-degree counted as d_0 (for instance 0),[18] when it is born at time t, then we can solve the differential equation given by (3) to find

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

where $r = \frac{p_r m_r}{p_s m_s}$ is the ratio of the number of links that are formed at random compared to through the search.[19]

THEOREM **1** *The 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 $d \geq d_0$.

The proof of Theorem 1 follows from Lemma 1, which appears in the appendix, and uses standard techniques of mean-field approximations.

As a check on the mean-field approximations, we match the analytic solution from Theorem 1 with simulations of the random process itself. The two match up well for all degrees, and for a variety of different parameters that we have run. Figure 1 shows typical comparisons. The red curves are the predictions from (5) while the black dots represent the empirical distributions derived from the simulation.[20]

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

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

[18]We allow this to potentially differ from 0, again so that we can compare this to other models, such as preferential attachment where it is necessary to start the in degree at a level different from 0, or a node would never get any links.

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

[20]Each panel depicts the results of a single (typical)computer simulation. We have not run enough simulations to estimate standard errors from the theoretical predictions.

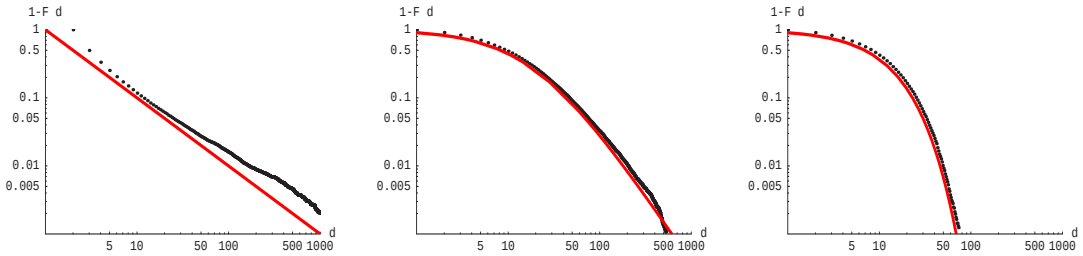

Figure 1. *Simulations are based on $T = 25,000$ periods. The red curve is the prediction from (5) while the black dots represent the empirical distribution derived from the simulation. (Left) All links are formed through search and the resulting distribution is scale free ($m_r = m_s = 10$, and $p_r = 0$, $p_s = 1$, starting each node with an in-degree of one to ensure that entering nodes can be found). (middle) Equal balance between search and random attachment, with a degree distribution that is scale-free in the upper but not lower tail ($m_r = m_s = 10$, and $p_r = p_s = 1$). (Right) All links are formed at random and the degree distribution is not scale-free in either tail ($m_r = m_s = 10$, $p_r = 1$, and $p_s = 0$).*

For large d relative to rm, (5) becomes linear in $log(d)$, and thus approximates a scale-free distribution. *However, for small d, the expression will not approximate a scale-free distribution.* This makes intuitive sense, as the extreme case where $r = 0$, links are only formed via search and the parents are never attached to. Then nodes are linked to exclusively proportionally to how easy they are to locate via search, and this corresponds to pure preferential attachment.[21] On the other extreme, when p_s and/or $m_s = 0$, the process is one of purely random link formation. Then (again, see the appendix for details),

$$\log(1 - F(d)) = \frac{d_0 - d}{b}, \tag{6}$$

which is a negative exponential distribution.[22]

Figure 1 illustrates these effects by showing the complimentary cdf $1 - F(d)$ of the degree distribution for three quite different parameterizations of the model. The left panel shows

[21] To have pure preferential attachment rather than search, (3) should be rewritten as $\frac{dd_i(t)}{dt} = \frac{md_i(t)}{tm+td_0}$, as the d_0 (which might be fictitious in nodes to make sure that nodes start with some degree, as in Bollobas and Riordan [11]) matters in the preferential count. The corresponding solution for the complementary cdf (see the appendix for omitted details) is then $1 - F(d) = d^{-\frac{m+d_0}{m}}$ which corresponds to the well known $Prob(d) \sim d^{-3}$ when new nodes have $d_0 = m$.

[22] Note that the fact that this is a growing system distinguishes it from a static random network, and so the degree distribution differs from what one would find in some of the networks analyzed by Erdös and Rényi [19].

a case where the roles of random and search linking are roughly balanced, and generates a degree distribution that is nearly scale-free. In contrast, the middle simulation shows a case where the majority of links are formed randomly. In this case the degree distribution is nearly scale-free in the upper tail, but the lower tail is distinctly thinner than a scale-free distribution would predict. The third case (right panel) is a purely random specification where no links are formed via search, and so the degree distribution is not scale-free in either tail.

Compare these distributions with those in Figure 2, which contains data from co-authorship networks (left) from Newman [48] and the world wide web from Albert, Jeong, and Barabási [3]. While the latter appears to approximate a scale-free distribution (we come back to a more precise fit in Section 3), the former clearly does not.

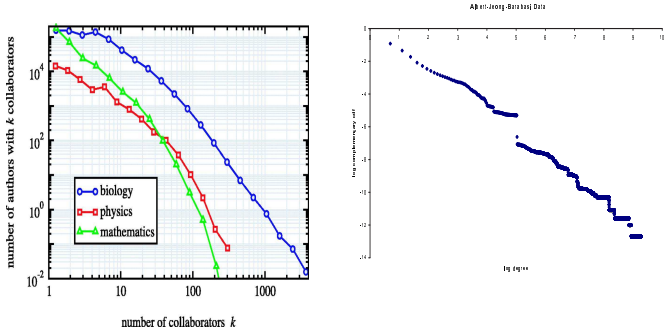

Figure 2. *(left) Data from Newman [48] containing the frequencies of authors with varying numbers of coauthors, which are clearly not scale free. (right) Data from Albert, Jeong, and Barabási [3] showing the complimentary cdf of web-page degrees.*

2.2 Clustering

We now turn to analyzing the clustering coefficient.[23] There are several ways in which one might measure clustering, and numerous variations appear in the literature. We examine three common measures.

The first is a well-known measure from the sociology literature (e.g., see Wasserman and Faust [57]) that examines the percentage of "transitive triples." This looks at situations

[23]We apologize for the use of the term clustering, which has other connotations in the sociology literature. We follow the terminology from the recent literature on large networks, in order to make some explicit comparisons.

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. 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).[24] That is, setting $\widehat{g}_{ij} = \max[g_{ij}, g_{ji}]$, we have an alternative measure of clustering where the directed nature of the links are ignored and we only pay attention to whether there is some relationship between nodes.[25] This 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 that is used is similar to the 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 [58], as well as many empirical studies (e.g., see the survey by Newman [47]).[26]

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

One might expect that these measures would be similar, but they can differ significantly,[27] as we shall see.

The analysis of some of the clustering coefficients requires knowledge of the degree distributions, and some involved conditional probability calculations. In order to obtain closed form solutions for these coefficients, we use mean- field approximations to simplify some

[24] See Newman [47].

[25] There are also hybrid measures (mixing ideas of directed and non-directed links) where one counts the percentage of possible directed links among a given node's direct neighbors that are present, on average, as examined by Adamic [1] on the www.

[26] One might ask why not also consider an individual average version of the fraction of transitive triples. One generally can. However, for our search model that measure coincides with the fraction of transitive triples.

[27] See Table II in Newman [47] for some illustration of the differences between $C(g)$ and $C^{Avg}(g)$.

expressions. 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 $\frac{p_r}{r}$ to be an integer and that exactly $\frac{p_r}{r}$ nodes are formed in each parent's neighborhood. (Recall that $r = \frac{p_r m_r}{p_s m_s}$, and so $\frac{p_r}{r} = \frac{p_s m_s}{m_r}$ represents the expected number of links to be formed through search per parent node identified.) This provides a tractable approximation to the more general process.

THEOREM **2** *Under a mean-field approximation to the model:*

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

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

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

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

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

$$\int_0^\infty \left[\frac{(rm)^{r+1}(r+1)}{(d+rm)^{r+2}}\right] \left(\frac{m^2 C^{TT}\left(1+\frac{2d(1+r)}{m}\right) - p_r d + rm\left[log\left(\frac{d}{rm}+1\right)\right]\left(\frac{p_r}{r}+p_r-2C^{TT}m(1+r)\right)}{(d+m)(d+m-1)/2}\right) dd$$

There are some interesting features to note about the various clustering coefficients.

First, in the special extremes of the model of a purely random network ($p_s = 0$), and the other extreme of a pure preferential attachment network ($p_r = 0$), all three coefficients are 0. As we alluded to earlier, neither of those models is a good fit of reality, as the data exhibits significant clustering. Indeed, substantial evidence has accumulated suggesting that large decentralized networks very generally exhibit clustering measures much larger than would be predicted by either purely random processes or models based on preferential attachment. For instance, Watts [58] gives an average clustering coefficient of 0.79 for the network consisting of movie actors linked by movies in which they have co-starred and Newman [47] reports a total clustering coefficient of 0.20 for the same network. Networks of researchers linked by co-authored papers have also been analyzed in various fields of study. Newman [48] gives total clustering coefficients of 0.496 for computer science, and 0.43 for physics, while Grossman [29] gives a measure of 0.15 in mathematics. Several authors have also analyzed clustering in the world wide web. For instance, Adamic [1] gives an average clustering measure of 0.1078 on a portion of the web containing over 150,000 sites (compared to 0.00023 for a purely random graph of the same order and number of edges).

Second, when both random attachment and search are present the fraction of transitive triples and the average clustering coefficient are bounded away from 0,[28] in contrast to the extremes of the model. The fact that these clustering coefficients do not vanish here comes from the combination of the random and search parts of the model. It is likely that a given node links to two different nodes who are linked to each other, precisely because they are linked to each other. This is the critical feature that distinguishes our search-based model from 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.[29] Previous models that have been shown to generate high clustering either start from some lattice structure and then rewire, as in Watts and Strogatz [59], involve some hierarchical structure (see Eiron and McCurley [18]), or, as in Klemm and Eguíluz [36],[37], require entering nodes to link to an entire population of active nodes that changes very slowly with time.

A final point is that the *total* clustering coefficient is nonzero only when $r > 1$, that is, only in cases where random attachment is more prevalent than search. The intuition for this is easily seen. One factor is that the variance of the degree distribution $F_t(d)$ (see Theorem 1) is finite only when $r > 1$. When $r < 1$, the predominance of links are formed through search and larger nodes grow at a fast enough rate so that the variance explodes. The relationship to clustering comes from the fact that nodes with huge in-degrees have very low clustering rates in that the many nodes that have connected to them are relatively unlikely to be linked to each other.[30] When these nodes with large in-degrees form a large enough fraction of the population, the total clustering coefficient tends to 0. This provides the contrast with the average clustering coefficient, since then the large nodes receive relatively little weight, while in the total clustering coefficient calculation nodes are essentially weighted by their degree.

We return in Section 3 to fit our model to the data.

[28]To see that the average clustering coefficient is bounded below, note that a lower bound on $m^2 C^{TT}$ is $p_r p_s m_s$. Then a lower bound on the integral is $\int_0^\infty f(d) \left(\frac{p_r p_s m_s + d p_r}{(d+m)(d+m-1)} \right) dd$, 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 $m \geq 1$.

[29]In such hybrid models the clustering coefficients also tend to 0, as for instance, shown by see Fronczak, Fronczak, and Holyst [23].

[30]The reason that this does not manifest itself in the fraction of transitive triples calculations is that these sorts of pairs of links (both pointing in to a given node) are not part of the relevant basis of that calculation.

2.3 Diameter

Diameter is difficult to establish in the context of a random graph, especially when the structure strays from the purely random structure first studied by Erdös and Rényi [19].

For some special cases we can deduce limits on the diameter by piggy-backing on powerful results due to Bollobás and Riordan [11]. In particular, they show that a preferential attachment network formation process where each node forms a single link (see also Reed [52]) 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:

THEOREM **3** *If $p_r = 0$, $p_s = 1$, $m_s = 1$, and $m_r \geq 2$, then the resulting network will consist of a single component with diameter*[31] *proportional to $\frac{\log(t)}{\log\log(t)}$, almost surely.*

The proof follows from Bollobás and Riordan [11].[32]

We conjecture that increasing the parameters p_r and m_s and decreasing p_s (provided $m_r \geq 2$) will not affect these results, as this simply leads to an increased number of links in the network. This is confirmed by following the heuristic test suggested on page 24 of [11]. However, once the parameters p_r and m_s are increased, the process is no longer covered by the [11] approach. Moreover, the system seems to be complicated enough so that no previous techniques for establishing tight limiting diameters apply. It is worth noting that the constraint that $m_r \geq 2$ is critical. It is this attachment to at least two independent neighborhoods that allows a node to form a bridge between different existing neighborhoods of the network, thus reducing path lengths. Moreover, the fact that the search method is more likely to lead to nodes with relatively very large degree means that new links are likely to lead to shortening paths between many existing nodes. In contrast, in a case where only one neighborhood is searched, this bridging no longer takes place and the diameter stays on the order of that of a purely random network ($\log(t)$).

Thus, 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.

[31] Given the directed nature of the links, diameter is measured based on paths where a link can go in either direction. Clearly, the diameter will generally be infinite if we measure paths in other directions, as some nodes will form no outward links whatsoever under the general random process we have described.

[32] 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.

2.4 Assortativity

As Newman [47] notes, a further feature distinguishing socially generated networks from other networks (e.g., 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 often referred to as assortativity, as this means that higher degree nodes have a greater relative tendency to be linked to each other, as in an assortative matching.

Generally, many growing models of networks will exhibit assortativity, as we also see assortativity in other models, e.g., by Krapivsky and Redner [38] as well as Callaway et al [12]. 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 older nodes tend to be connected to older nodes, and this means that nodes with higher degree are relatively more likely to be connected to each other. This leads to a positive correlation among the degree of connected nodes in the network.

This can be proven formally as follows. 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.[33] 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 search), if* $d_i(t) > d_j(t)$, *then* $1 - F_i^t(d) > 1 - F_j^t(d)$ *for all* $d < d_i(t)$.[34]

This result provides a stronger relationship than just noting a positive correlation, as it establishes a form of first-order stochastic dominance of the degree distribution of a node's neighbors.

2.5 Negative Clustering-Degree Relationship

The clustering coefficient of a given node is inversely related to its degree in many socially generated networks. That is, if one examines a high degree node and asks what the probability is that a randomly chosen pair of its neighbors are linked, the answer is lower than the corresponding question for a low degree node. This, for instance, has been documented in co-authorship networks (e.g., see Goyal van der Leij, and Moraga-González [28]) among others. We refer to this as a *negative clustering-degree* relationship.

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

[34] 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.

The intuition of why the search model exhibits a negative clustering-degree relationship is the following. First, there is the effect that we mentioned before: as a node's neighborhood grows, the number of potential pairs increases quadratically, but the number of links increases only linearly. Second, there are some specific characteristics of the formation process that influence the clustering as a function of degree. Any particular node i gains links from nodes which find i either at random or through search of the network. Nodes that find i randomly contribute more to the clustering in i's neighborhood, since, after finding i, they search i's neighborhood and potentially connect to several of i's neighbors. As a node's in-degree grows, it becomes relatively less likely to gain additional links through random meetings relative to search, so that the clustering in it's neighborhood decreases.

This negative clustering-degree relationship is the reason why the total clustering coefficient $C(g)$ tends to zero (when $r < 1$, i.e., the role of random meeting is smaller than that of search), while the average clustering coefficient $C^{Avg}(g)$ remains bounded away from zero. Nodes with the largest degree have the smallest clustering in their neighborhoods, but receive a disproportionately low weight in the calculation of the average clustering coefficient. In contrast, the total clustering coefficient effectively weights nodes by their degree, so that the overall clustering measure is driven to zero.

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 yet have a formal proof. Nonetheless, it is quite easy to prove a weaker version of this statement. In particular, for nodes of sufficiently large degree, if two nodes have a large enough difference in their degrees, then the node with larger degree has a smaller clustering coefficient. We make this precise in the following way.

THEOREM **5** *Under the mean-field approximation to the model (with nontrivial search), there exists $\bar{d} > 0$ such that for all $d > \bar{d}$ there exists $D > 0$ so that for all $d' > d+D$, $C(d) > C(d')$.*

3 Fitting the Model to Data

To demonstrate the power and flexibility of the model and some of the things that one can learn from it, we fit it to six data sets 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,[35] a citation network of research articles stemming from Milgram's [45] 1960 paper (all papers either reference Milgram [45] 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.[36] We show that the characteristics predicted by the model closely match the observed characteristics of these networks.

We fit the data as follows. First, we can directly calculate average in-degree and obtain m. With m in hand, and setting $d_0 = 0$, the only free parameter in the degree distribution F from (4) is r. Thus, by fitting the observed degree distribution we obtain an estimate of r. As the form for F in (4) is nonlinear, entering in a form $(d + rm)^{-(1+r)}$, we use an iterative procedure where we start with an initial guess for r, say $\widehat{r}$ and then plug this in to get an expression of the form $(d + \widehat{r}m)^{-(1+\widehat{r})}$. Using OLS (after taking logs) we then estimate r. We then iterate this process until we find a fixed point.[37] Next, we fit the clustering. We constrain $p_s = p_r = p$, as this seems a reasonable starting assumption and it eliminates a degree of freedom.[38] Then using our expressions for clustering we can estimate p. This ties down the parameters of the model. To estimate diameter, we then run simulations of the model based on the estimated m, r, and p, with the appropriate number of nodes. From the simulations, we obtain an estimated diameter which we can compare with the data.[39]

We first describe the fit to the www data. The www data has an average in-degree of about $m = 4.5$. We fit the degree distribution from our model to the data to find a ratio $r = .5$ of random to local search in the process. The fit is pictured below.

One thing that we again emphasize is that even though the model is clearly not scale-free overall, it looks quite linear in the above plot (left panel) and matches the data quite well

[35]This network, as well as that of the ham radio calls, is not directed, and so our model needs some modifications to apply (see the discussion in Section 5.2 for more discussion). In particular, p_r and p_s can be interpreted as the probability that both parties involved find it worthwhile to co-author together. Second, to strictly apply the model, we keep track of who found whom in the network. The other networks are directed.

[36]We obtained the www data from Albert, Jeong, and Barabási [3], the co-authorship data from by Goyal, van der Leij, and Moraga-González [28], the citation network from Garfield [26], the prison data from MacRae [44], the ham radio data from Killworth and Bernard [34], and the high school romance data from Bearman, Moody, and Sovel [10].

[37]Given this functional form, this iteration has nice monotonicity properties, converging to the same fixed point whether we start from very high or very low guesses of the initial $\widehat{r}$.

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

[39]We do not fit assortativity or clustering-degree relationships for these data sets, as we do not have precise measures for these either for the model or the data sets.

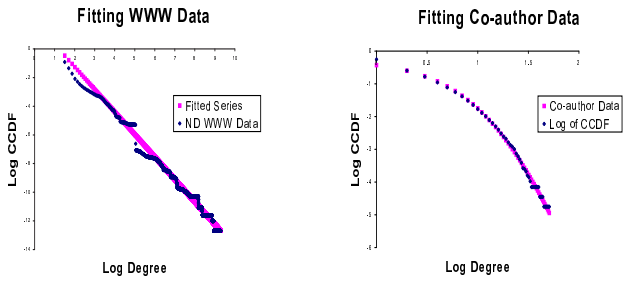

Figure 3. *Pink: Fit from our model; Blue: Data ([3], [28]); (left) The complimentary cdf of web page degrees, (right) The complimentary cdf of co-author degrees.*

(with an R^2 of .97).[40]

The total clustering coefficient is then 0, while the average clustering coefficient is $.34p$ (where we work under the assumption that $\frac{p}{r} < 1$, given that $r = .5$). If p is roughly 1/3, then we match figure of 0.11 reported in Adamic [1] for his (different) www data set.[41]

In terms of the diameter, we only have an order of magnitude calculation from Theorem 3. Given that the number of nodes in this data set is almost $T = 326,000$, $\ln(T) = 12.7$ and $\frac{\ln(T)}{\ln\ln(T)} = 5.0$. Thus simply knowing an order of magnitude calculation does not even allow one to differentiate between a completely random network and one coming from the model, so we resort to simulations. We bound the diameter between 16 and 32 for $T = 326,000$ based on our simulations.[42] Our data includes only the degree distribution for the [3] www data, but Newman [47] reports an *average* distance of 11.27 for those data.

[40] If one fits the data linearly, rather than through our model, then one obtains $f(d) \sim d^{-2.56}$ (the corresponding pdf with a coefficient on the cdf of -1.56). This differs slightly from the figure of -2.1 reported by the authors, who fit the data after coalescing the data into bins, rather than working with the data directly.

[41] We do not have a clustering coefficient for the Albert, Jeong, and Barabási [3] data. Newman [47] reports a figure of .29, referencing Albert, Jeong, and Barabási [3] and Barabási, Albert, Jeong [9], 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.

[42] 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 bounds of $\{11, 22\}$, $\{13, 26\}$, and $\{14, 28\}$, 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 of nodes). Thus we obtain upper and 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.

We perform a similar analysis for the co-authorship network of Goyal, van der Leij, and Moraga-González [28], with the data for the 1990's. The data has just under two links per researcher in a network with 81,217 researchers. Given the directed nature of our process, we simply adapt our model by keeping track of which nodes initiate the links, and now p simply represents the probability that both researchers find a link worthwhile.[43] Given that there are roughly two links per researcher, each researcher initiates one link on average, and receives one link initiated by another researcher on average. Thus, we set $m = 1$. Based on this, we estimate $r = 3.54$ with provides a fit with an R^2 of .99. This suggests that about three and one half times more links are formed at random compared to through local search. The fit is pictured in the right panel of Figure 3.

The predicted average clustering coefficient based on these estimates is $.94p$, which compares well with the reported figure of 0.16 for the community of economics researchers in the 1990s if we set $p = .17$. The total clustering coefficient is $.17p$. As for the diameter, simulations suggest that the diameter is bounded between 18 and 36 on a network of size $T = 81,217$, where the diameter is for the largest component if there is more than one component. The data has a diameter of 26, according to Goyal, van der Leij, and Moraga-González [28], again based on the largest component. In addition, our simulations predict the size of the giant component to be roughly 50,000, and the empirical size is 33,027.

In the citation network there is an average of 5 references per paper, and there are 396 papers in the data set. Setting $m = 5$, we estimate a value of $r = .62$ with an R^2 of .98. The fit is pictured in Figure 4 (bottom left panel). Based on the estimates of m and r, we compute the average clustering coefficient to be $.26p$. When $p = .26$ this matches the empirical figure of .067. The diameter of this network is 4, which is consistent with our simulations, which result in diameters of either 4 or 5.

The 67 prison inmates named an average of $m = 2.7$ friends. Performing the same exercise, we estimate $r = 590$ with an R^2 of .94.[44] The data and fit are depicted in Figure 4 (top right panel). Average clustering is $.0012p$. In order to match the actual figure of .0011 we require p to be nearly unity. Simulations suggest a diameter of 5, while the actual figure for this data set is 7.

[43] Some papers in the 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 their data. 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.

[44] While one might think of this as a non-directed network, the data are actually directed, as it contains the reports of whom a given inmate considers to be a friend, which is not always a reciprocal relationship. The same is true of the high-school romance network fitted below.

In the ham radio network there is an average of 3.5 links per ham radio operator, with 44 operators in the data set.[45] We estimate a value of $r = 5$ with an R^2 of .94. The fit is pictured in Figure 4 (bottom left panel). Based on the estimates of m and r, we compute the average clustering coefficient to be $.09p$. In the data, average clustering is .06, while total clustering is .20. We note that total clustering is greater in this data, an anomaly we attribute to the small size.

In the high school romance network of Bearman et al [10] there is an average of .84 partners per student, with 572 students in the data set. Setting $m = .84$, we estimate a value of $r = 1000$ with an R^2 of .99. The fit is pictured in Figure 5 (left panel). Given that this network is largely heterosexual, there is essentially no clustering in the data, and so we do not fit that, and we do not have the diameter information for this network.[46]

One of the interesting things that comes out of this analysis is that we can compare the relative ratios of the number of nodes located at random relative to search across the two data sets. Table 1 summarizes these findings. The search model suggests, for example, that the role of local search is much more prevalent in the formation process of www network than for the co-author network. Specifically, the random-to-search ratio is approximately 1/2 in the www data and 3.5/1 in the economics co-author network. Thus local search is seven times more prevalent in the www network formation process than in the formation of the co-author network.

Table 1: Comparing r Across Applications

	WWW	Citations	Co-author	Ham Radio	Prison	High School Romance
r	0.5	0.62	3.5	5.0	590	1000

4 Efficiency and Network Structure

While it is of direct interest to have estimates of r, and to see that networks differ extensively in terms of the underlying formation processes; it is also of interest to understand the

[45] This is also a non-directed network, and so we fit the model as we did for the co- authorship data.

[46] A proper modeling of this network would require allowing for different sexes and for most nodes to only link to nodes of the opposite sex. While we do not pursue that here, it would be any easy extension of our model. The degree distribution is still matched quite accurately by our model in spite of this deficiency.

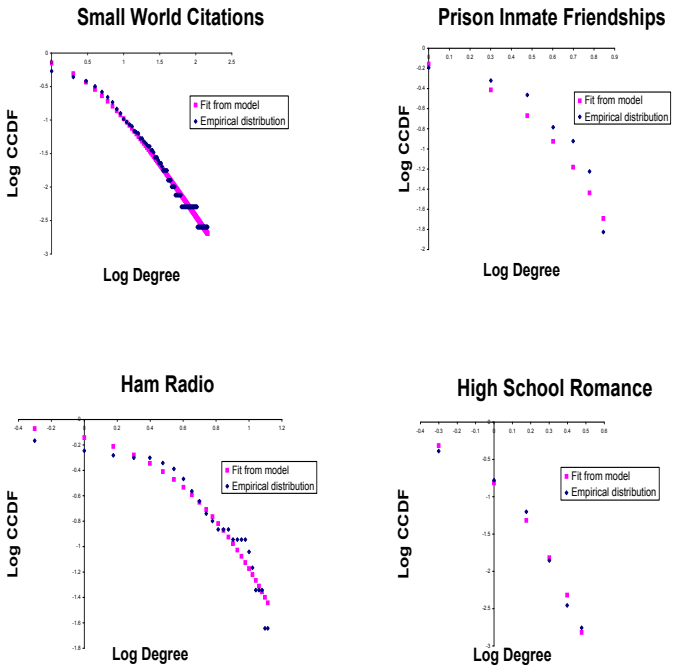

Figure 4. *Pink: Fit from our model; Blue: Data ([26], [44]); (top left) The complimentary cdf of the degree sequence in the citation network, (top right) The complimentary cdf of the prison inmate network, (bottom left) The complimentary cdf of the degree sequence in the ham-radio network, (bottom right) The complimentary cdf of the high school romance network.*

implications of r for the operation of a network. That is, we would like to know whether a high or low r is a "good" or "bad" thing in terms of the function of a network. Based on the model, we can actually say quite a bit about this. A very helpful result is that as we vary r, the resulting degree distributions can be completely ordered in terms of strictly second order stochastic dominance.

4.1 Degree Distributions and Second Order Stochastic Dominance

THEOREM **6** *Consider the distribution function described in (4) and any fixed $m > 0$ and set $d_0 = 0$.*[47] *allowing for values If $r' > r$, then F' strictly second order stochastic dominates*

[47]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

F, where F' and F are the distribution functions corresponding to r' and r, respectively.

Theorem 6 shows that we can completely order the degree distributions that emerge as we vary the percentage of nodes that are formed at random versus through local search, in the sense of second order stochastic dominance. This 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 in terms of r. Any model where there are diminishing marginal utilities to adding additional links to a given node will satisfy this.

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). Then for any given m and d_0, 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.*[48]

In the above analysis we have varied r while holding m fixed. One can also do the reverse, to find that higher values of m correspond to degree distributions that *first* order stochastic dominate degree distributions with lower m, which 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 of this theorem, as it can be 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 function of the node's degree, there are also some important contexts where we cannot express things in this manner. For instance, in some cases the degrees of the nodes that are connected to a given node are also important in determining the given node's utility. Such situations are more difficult to order in terms of overall welfare. Nevertheless, we can still say something even in those cases. Let us describe one such application that is of interest.

$d_0 = 0$ is the relevant one.

[48] 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.

4.2 Infection Rates and Network Structure

Consider the spread of a virus, disease, or even computer virus, through a network. One can also think of applications to the spread of behavior or information. A standard model of such spreading is the SIS model (susceptible, infected, susceptible model, see Bailey [5]). Here we follow recent analyses by Pastor-Satorras and Vespignani [50] and Lopez-Pintado [42], that allow one to estimate infection rates based on degree distributions. We adapt those models to our setting as follows.

Consider a network where a given healthy node catches a disease in a given period with a probability $\nu d_i \rho_i$, where $\nu \in (0,1)$ is a parameter describing a rate of transmission of infection in a given period, d_i is the (in-)degree of i, and ρ_i is the probability that any given neighbor of i is infected.[49] Also suppose that any infected node recovers in a given period with a probability $\delta \in (0,1)$. We can then ask how the long-run steady-state proportion of infected nodes relates to the network structure. Using mean-field approximations, we derive an ordering of infection rates based on the parameters of the degree distributions in (4).

We should emphasize that the following result uses an assumption that the degree distribution across nodes is independent. As shown in Theorem 4, this is inconsistent with our model. Thus, the theorem below does not apply to our model, but only to networks having a degree distribution in the family that we derived from our model. Analyzing the relationship with significant correlation among nodes appears to be a difficult problem. Despite the mismatch of Proposition 1 with the correlation details of the model, we still feel it is of enough interest to present.

PROPOSITION **1** *Consider a network with a degree distribution described by (4) for a given m and d_0 which is independent across nodes. For any r and r', with $r' > r$, there exist $\underline{\lambda}$ and $\overline{\lambda}$ such that*

- *If $\frac{\nu}{\delta} < \underline{\lambda}$ then the steady-state average infection rate under a mean-field approximation is lower under r' than r.*
- *If $\frac{\nu}{\delta} > \overline{\lambda}$ then the steady-state average infection rate under a mean-field approximation is higher under r' than r.*

The theorem again uses mean-field approximations, where average infection rates among nodes of a given degree are approximated by a continuous time process. Steady-states are

[49] For small ν, this is an approximation of becoming infected independently from any of i's infected neighbors.

found by setting the change in average infection rates over time to 0. For any given degree distribution, there may exist multiple steady-state infection rates. For instance, 0 is always a steady-state infection rate. We consider the largest steady-state infection rate when there are multiple infection rates.[50]

The intuition behind these results can be expressed as follows. The change in infection rate due to a change in the degree distribution comes from countervailing sources, as more extreme distributions have relatively more very high degree nodes and very low degree nodes. Very high degree nodes have high infection rates and serve as conduits for infection, thus putting upward pressure on average infection. Very low degree nodes have fewer neighbors to become infected by and thus have relatively low infection rates. Which of these two forces is the more important one depends on the ratio $\frac{\nu}{\delta}$. For low $\frac{\nu}{\delta}$, the first effect is the more important one, as nodes recover relatively rapidly, and so there must be nodes with many neighbors in order keep the infection from dying out. In contrast, when $\frac{\nu}{\delta}$ is high, then nodes become infected more quickly than they recover. Here the more important effect is the second one, as most nodes tend to have high infection rates, and so how many neighbors a given node has is more important than how well those neighbors are connected.

5 Concluding Discussion and Extensions

We have presented a model of network formation that exhibits features that match observed networks. As the parameters of the model are varied, the emerging networks change in specific ways that allow us to fit data and to derive results concerning efficiency.

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. The good news is that the model fits data remarkably well, and that we can derive

[50]These approximations do have some differences from the finite node system, which over time will eventually hit an absorbing state of 0 infection. In order to have an appropriate approximation, the finite system needs to have some random perturbations so that periodically fresh infections arrive if the system happens to hit the 0-infection state.

implications of the process for welfare and other characteristics. Nevertheless, it would be nice to delve deeper into the micro details of the link formation. With that in mind, let us discuss how the model extends along various dimensions.

5.1 Degree-Dependent Utility and Externalities

One important dimension along which to consider enriching the model is in terms of the utility structure. In the model discussed up to this point, 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, we can think of various reasons why the utility might actually be network- and degree-dependent. It might be that a given node enjoys benefits from indirect connections, and thus might be more likely to be willing to link to nodes that have larger degrees.[51] Alternatively, there might be correlation in the valuations across different nodes, and so higher degree might be related to a higher expected valuation for a given node.

Let us examine a simple variation on the model where the utility from attaching to a given node is proportional to its degree. In particular, suppose that the marginal utility obtained from linking to a node j is

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

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 the node 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 search process is proportional to $1 - \frac{c}{ud_j}$ (assuming that $d_j \geq 1$).

Let us consider a heuristic argument to see how this might affect the degree distribution. To keep things simple, let us suppose that parent nodes are attached to with certainty and that this utility calculation only enters for nodes identified through neighborhood search. We then end up with a mean-field process governed by

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

[51] An example of a 'connections' model where utility is derived from indirect connections was studied by Jackson and Wolinsky [32]. (See Jackson [30] for a survey of the related literature.) Recent variations on the connections model (where there is no decay across links - so that only shortest paths matter) have been analyzed in the context of large networks by Fabrikant et al [21] and Chun et al [15]. However, those analyses do not shed light on the issues studied here.

or

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

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), we find that

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

While the parameters have changed, the basic expression is similar to what we found before. Essentially, this tilts things more towards a preferential attachment model - so that the degree distribution looks approximately scale free at a lower degree.

While arguably many applications are captured by a model with either direct utility per node, or else well-approximated by something proportional to degree, there are also important applications (for instance, the passing of information) where the fuller network structure matters. Future research should investigate how other utility formulations impact the network formation process.

5.2 Non-directed Networks

As we mentioned, the analysis above extends to the case of non-directed networks with relatively minor variations. One change is that in the case of non-directed or bilateral relationships, we should generally expect that the consent of both parties is needed in order to form the link. Thus, the link will be formed if and only if both nodes get positive net utility from the interaction. This modification of the above analysis is a trivial one, as it simply reinterprets the parameters p_r and p_s as the probability that *both* nodes find the link attractive. There are also two possible variations in terms of how to adapt the search process. A straightforward variation is to implicitly keep track of who initiates a link, and to then apply our model directly. A more complicated variation is to ignore such information, so that search occurs through all of a parent's links, rather than just the ones they initiated. This leads to complications as now parent nodes will have degrees that are correlated with their age and that of their connections.[52] This mitigates the degree-dependence of the attachment process, as large degree nodes are more likely to be connected to each other, simply due to

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

their age. As a result, since a given node is more likely to be found when the parent node that it is connected to has fewer links, this provides a counter-bias to the benefit of having a high degree. As this bias grows relatively slowly in t, for large degrees one should still approximate a scale free distribution, but the specific details of the distribution could change from the directed case.

5.3 More Extensive Search

Suppose that we alter the model so that search extends uniformly over (directed) neighborhoods of path length greater than 1 from the parent node. This would lead to some slight adjustments of the formulas we examined before. First, the clustering coefficients would be lower, but still bounded away from 0. Second, 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. This biases things a bit more towards random attachment and away from preferential attachment, as the probability of being found via search is decreased (proportionally) by the size of the increase in neighborhood size).

5.4 Exponential Growth

The process described above has a single node born at each point in time. In some cases, the system will actually be growing exponentially. Modeling exponential growth does influence the degree distribution. To see how, let us examine an extension of the model to a growing number of nodes entering at each date - say proportionally to population size. 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

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

As shown in Lemma 3 in the appendix, this results in a complementary cdf of

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

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

5.5 Out-degree and Search by Existing Nodes

One dimension along which our model is clearly too restrictive is that all nodes have (roughly) the same out-degree. One easy extension would be to allow for random m_r and m_s, or else heterogeneous p_r and p_s across nodes. More generally, in many network applications, it is not simply new nodes that are forming new links, but links evolve on a constant basis. Adding search by existing nodes is easily incorporated to our model, by having some existing nodes search over time. The main complications in either of these 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 future analysis.

5.6 Other Applications

Although we have interpreted our model as a search-based 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 things like city size.[53]

The data in the figure below represent the population sizes of all counties in the US.[54] The upper tail of the distribution is roughly linear which is consistent with the fact that the literature has claimed scale-free distributions for city populations.[55]

It is important to remark, however, that the graph has a noticeable bend to it. Again, this is an important point to emphasize regarding power laws and log-log plots, which although very basic is easy to overlook. On 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.[56] Thus, when one plots degree distribution and other distributions, and simply fits a line and concludes that things are scale-free, one must be very careful in taking that conclusion seriously.[57] Indeed, it might be only true for relatively large measurements that this even approximately holds; i.e., only for the upper tail.

[53]This is often known as Zipf's law [61], even though Zipf was concerned with many things including word usage. See Gabaix [24] for a recent model of city growth and a discussion of Zipf's law.

[54]We thank David Alderson for sharing these data with us!

[55]We remark that the literature often focuses on the largest areas in terms of populations (for instance, Gabaix's [24] 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.

[56]Gell-Mann [27] suggests a function like the one proposed here would better match the data.

[57]Again, see Mitzenmacher [45] for more discussion of this, and the similarity between Pareto and lognormal distributions, for instance.

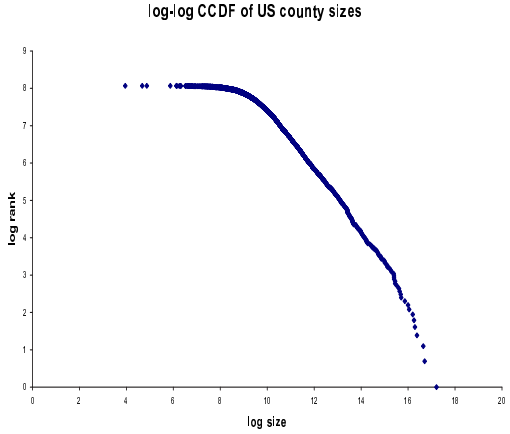

Figure 6. *log-log plot of the complimentary cdf of US county sizes showing that while the tail of the distribution is approximately scale free, the remainder of the data (over 90%) is not.*

This pattern of having the upper tail be scale-free, but the lower tail flatten out is exactly consistent with the model we have presented here. To understand how our model relates 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 to be close to a particular job location. Only slight variations in the model presented above need to be made in order to accommodate this setting.

References

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

[2] Albert, R., and A. Barabási (2002), "Statistical mechanics of complex networks," *Reviews of Modern Physics,* **74**: 47-97.

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

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

[5] Bailey, N.T.J. (1975) *The Mathematical Theory of Infectious Diseases*, Griffin: London.

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

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

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

[9] 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.

[10] Bearman, P., J. Moody, and K. Stovel (2004), "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks," Manuscript, University of Chicago.

[11] Bollobás, B., and O. Riordan (2002), "The diameter of a scale-free random graph,", Manuscript, to appear.

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

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

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

[15] Chun, B-G., R. Fonseca, I. Stoica, and J. Kubiatowicz (2004) "Characterizing Selfishly Constructed Overlay Routing Networks," *Proceedings of the 23rd IEEE International Conference on Computer Communications (INFOCOMM).*

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

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

[18] Eiron, N. and K.S. McCurley (2003) "Locality, Hierarchy, and Bidirectionality in the Web," Extended Abstract for the sl WAW 2003.

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

[20] Fabrikant, A., E. Koutsoupias, and C. 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.*

[21] Fabrikant, A., A. Luthra, E. Maneva, C. Papadimitriou, and S. Shenker (2004) "On a Network Creation Game," preprint: U.C. Berkeley.

[22] Faloutsos, M., P. Faloutsos, and C. Faloutsos (2004) "On Power-Law Relationships of the Internet Topology," preprint: U.C. Riverside.

[23] Fronczak, A., P. Fronczak, and J.A. Holyst (2003) "Mean-Field Theory for Clustering Coefficients in Barabási-Albert Networks," arXiv:cond- math/0306255 v1, 10 June.

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

[25] Galeotti, A., S. Goyal, and J. Kamphorst (2004) "Network formation with heterogeneous players," preprint: Tinbergen Institute.

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

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

[28] Goyal, S., M. van der Leij, and J. L. Moraga-González (2003). "Economics: an emerging small world?," preprint: University of Essex.

[29] Grossman, J. 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).

[30] Jackson, M.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., http://www.hss.caltech.edu/∼jacksonm/netsurv.pdf.

[31] Jackson, M.O. and B. Rogers (2004) "The Economics of Small Worlds," forthcoming in the *Journal of the European Economic Association - Papers and Proceedings)*, http://www.hss.caltech.edu/∼jacksonm/netsmall.pdf.

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

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

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

[35] Kleinberg, J.M., S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. 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, K. and V.M. Eguíluz (2002) "Growing Scale-Free Networks with Small World Behavior," Physical Review E, vol 65(3), 036123,

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

[38] Krapivsky, P.L. and S. 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, E. Upfal (2000) "Stochastic Models for the Web Graph" *FOCS 2000.*

[40] Levene, M., T. Fenner, G. Loizou, and R. Wheeldon (2002) "A Stochastic Model for the Evolution of the Web," *Computer Networks*, vol 39: 277-287.

[41] Li, L, D. Alderson, W. Willinger, J. Doyle, R. Tanaka, and S. Low (2004) "A First Principles Approach to Understanding the Internet's Router Technology," *Proc. Sigcomm*, ACM.

[42] Lopez-Pintado, D. (2004) "Diffusion in Complex Social Networks," Universidad de Alicante WP-AD 2004-33.

[43] Lotka, A.J. (1926) "The Frequency distribution of scientific productivity," *Journal of the Washington Academy of Sciences*, Vol. 16, 317-323.

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

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

[46] Mitzenmacher, M. "A Brief History of Generative Models for Power Law and Lognormal Distributions.", Manuscript. http://www.eecs.harvard.edu/∼michaelm/ListByYear.html.

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

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

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

[50] Pastor-Satorras, R. and A. Vespignani (2000) "Epidemic Spreading in Scale-Free Networks," *Physical Review Letters*, 86:14, 3200-3203.

[51] Pennock, D.M., G.W. Flake, S. Lawrence, E.J. Glover, and C.L. Giles (2002) "Winners don't take all: Characterizing the competition for links on the web," *PNAS*, 99:8, pp. 5207-5211.

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

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

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

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

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

[57] Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press.

[58] Watts, D. (1999), "Small Worlds," Princeton University Press.

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

[60] Yule, G. (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.

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

6 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

$$1 - F_t(d) = \left(\frac{d_0 + \frac{b}{a} - \frac{ct}{1-a}}{d + \frac{b}{a} - \frac{ct}{1-a}} \right)^{1/a}. \tag{7}$$

If $a = 0$ and $c = 0$, then (9) has solution

$$1 - F_t(d) = e^{\frac{d_0 - d}{b}}. \tag{8}$$

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

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

$$\frac{dd(t)}{dt} = \frac{ad(t)}{t} + \frac{b}{t} + c, \tag{9}$$

with initial condition $d(i) = d_0$ (where $i < t$). If $a > 0$ and either $c = 0$ or $a \neq 1$, then (9) 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 (9) 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

$$d_i(t) = \left(d_0 + \frac{b}{a} - \frac{ct}{1-a}\right)\left(\frac{t}{i}\right)^a - \frac{b}{a} + \frac{ct}{1-a}. \tag{10}$$

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

$$d_i(t) = d_0 + b\log\left(\frac{t}{i}\right) \tag{11}$$

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 (10) and (11) we deduce that

$$i^*(d) = t\left(\frac{d_0 + \frac{b}{a} - \frac{ct}{1-a}}{d + \frac{b}{a} - \frac{ct}{1-a}}\right)^{\frac{1}{a}}. \tag{12}$$

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

$$i^*(d) = te^{\frac{d_0 - d}{b}} \tag{13}$$

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

$$1 - F_t(d) = \left(\frac{d_0 + \frac{b}{a}}{d + \frac{b}{a}}\right)^{\log(1+g)/a}. \tag{14}$$

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 neighborhood search.

3. Both j and k were found by neighborhood search.

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

In case 2, we have $p_r(p_s m_s)$ such situations where the link to the random node was formed and then a link to a node in its neighborhood was formed; where $p_s m_s$ is the expected number of nodes formed through search 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 $\frac{p_s m_s}{m_r} = \frac{p_r}{r}$ links formed by a new node to one of its parent's neighborhoods and m_r parents, and so there are $m_r \frac{p_s m_s}{m_r}(\frac{p_s m_s}{m_r} - 1)/2$ 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

$$\frac{C^{TT}mp_sm_s}{m-1}\left(\frac{p_r}{r}-1\right) \tag{15}$$

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

$$p_rp_sm_s + \frac{C^{TT}mp_sm_s}{m-1}\left(\frac{p_sm_s}{m_r}-1\right)$$

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

$$C^{TT} = \frac{p_rp_sm_s + \frac{C^{TT}mp_sm_s}{m-1}\left(\frac{p_sm_s}{m_r}-1\right)}{m^2},$$

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

$$C^{TT} = \frac{p_rp_sm_s}{m^2}$$

otherwise. Solving for C^{TT} in the first equation yields the claimed expression.[59]

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^2C^{TT} such triples per node; for a total of tm^2C^{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

$$C = \frac{3m^2C^{TT}}{m(m-1)/2 + m^2 + \frac{1}{t}\sum_i d_i(d_i-1)/2}. \tag{16}$$

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

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

and a corresponding density function of

$$f_t(d) = (rm)^{r+1}(r+1)(d+rm)^{-r-2}. \tag{17}$$

[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.

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 (16) (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 (17) average clustering, $C^{Avg}(g)$, tends to

$$\int_0^\infty (rm)^{r+1} (r+1) (d + rm)^{-r-2} C(d) dd, \tag{18}$$

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

$$\frac{(d+m)(d+m-1)}{2} \tag{19}$$

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

$$C^{TT} m^2 \tag{20}$$

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 (11) tells us that

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

Then from equation (12), from the process of $d_i(t)$ (not to be confused with $d_i^r(t)$), it follows that

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

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 search has $p_s m_s/m_r$ links that are attached based on a search of i's neighborhood, and thus we have

$$rm\left[\log\left(\frac{d}{rm}+1\right)\right]\frac{p_s m_s}{m_r} \tag{21}$$

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 through search 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 search. 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 search. We can simply add the expected number of this type of completed triple over each node j that has attached to i through search. 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 search and also to some other node k through search. A fraction $\frac{2}{p_s m_s}$ of the potential pairs of outward links from j that are both formed through search 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

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

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

$$\left(d - rm\left[\log\left(\frac{d}{rm}+1\right)\right]\right)\left(p_r + \frac{2}{p_s m_s}\left(C^{TT}m^2 - p_r p_s m_s\right)\right) \tag{22}$$

for the number of completed triples of this type.

Finally, by summing (20), (21), and (22), we find the numerator of $C(d)$, and (19) provides the denominator. Plugging this expression for $C(d)$ into (18) 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 (10). ∎

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

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

Thus $C(d)$ is approximated by $\frac{\left(\frac{2m^2 C^{TT}}{p_s m_s} - p_r\right)d}{\frac{1}{2}d^2}$ for large d. Since this expression is decreasing in d, the result follows directly. ∎

Proof of Theorem 6: By standard results on second order stochastic dominance (e.g., Rothschild and Stiglitz [53]), it is sufficient to show that

$$\int_{d_0}^{X} [F(d) - F'(d)]dd > 0 \tag{23}$$

for all $X > 0$. Substituting from (4), we rewrite (23) 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 $X > 0$. ∎

Proof of Proposition 1: Let $\lambda = \nu/\delta$. Also let $\rho(d)$ denote the steady-state infection rate of a node with degree d, and ρ be the average across nodes: $\rho = \int \rho(d)dF(d)$. Let

$$\theta = \frac{\int d\rho(d)dF(d)}{\int d\ dF(d)} = \frac{\int d\rho(d)dF(d)}{m}.$$

In steady-state

$$\rho(d) = \frac{\lambda\theta d}{1 + \lambda\theta d}. \tag{24}$$

Multiplying both sides by $1 + \lambda\theta d$ and integrating with respect to $dF(d)$, we obtain

$$\rho + \lambda\theta^2 m = \lambda\theta m$$

or

$$\rho = \lambda\theta m(1 - \theta). \tag{25}$$

We use the following lemma

LEMMA 4 *If F' strictly second order stochastic dominates F, then the corresponding $\theta' < \theta$.*

Proof of Lemma 4: Multiply both sides of (24) by d and integrate with respect to $dF(d)$ to obtain

$$\theta = \int [\lambda\theta d^2/(1 + \lambda\theta d)]dF(d)/m. \tag{26}$$

Now consider F' that strictly second order stochastic dominates F, and consider any corresponding θ' and θ. Let us show that $\theta' < \theta$.

Suppose to the contrary that $\theta' > \theta$.

Since θ is the largest point in [0,1] such that

$$\theta = \int [\lambda d^2\theta/(1 + \lambda d\theta)]dF(d)/m,$$

it follows that

$$\theta' \neq \int [\lambda d^2\theta'/(1 + \lambda d\theta')]dF(d)/m.$$

Note that

$$\int [\lambda d^2/(1 + \lambda d)]dF(d)/m < \int [\lambda d^2/(\lambda d)]dF(d)/m = 1.$$

Thus, as x ranges from θ to 1,

$$\int [\lambda d^2 x/(1 + \lambda dx)]dF(d)/m$$

ranges from θ to something smaller than 1. Thus, for all $x > \theta$, it must be that

$$x > \int [\lambda d^2 x/(1 + \lambda dx)] dF(d)/m$$

as otherwise there would exist another fixed point since the left hand side ranges continuously from θ to 1, while the right hand side ranges continuously from θ to something smaller than 1. Therefore, since $\theta' > \theta$, it follows that

$$\theta' > \int [\lambda d^2 \theta'/(1 + \lambda d\theta')] dF(d)/m.$$

However, since $\lambda d^2\theta/(1 + \lambda d\theta)$ is strictly convex in d, it follows from the strict second order stochastic dominance of F' over F that

$$\theta' = \int [\lambda d^2 \theta'/(1 + \lambda d\theta')] dF'(d)/m < \int [\lambda d^2 \theta'/(1 + \lambda d\theta')] dF(d)/m.$$

We have reached a contradiction. Thus, $\theta' < \theta$ whenever F' strictly second order stochastic dominates F. ∎

From (25), we know that ρ is increasing in θ when θ is below 1/2; but decreasing when it is above 1/2.

From (26), we know that θ is near zero for low λ, and near one for large enough λ, for any given F. Given r and r', we can then find a bound on λ below which both θ and θ' are below 1/2, and corresponding a bound on λ above which both θ and θ' are above 1/2. The proposition then follows from Lemma 4 and (25). ∎