# Search and the Strategic Formation of Large Networks: When and Why do We See Power Laws and Small Worlds?

Matthew O. Jackson and Brian W. Rogers*

Extended Abstract for the P2P Conference
May 25, 2004

## 1 Introduction

Network structures play a central role in determining outcomes in many important situations. Examples include the world wide web, joint research venture projects among firms, co-author relationships among academics, political alliances, trade networks, the organization of intra-firm management, social networks for transmitting information, and P2P systems for file sharing. Given the large and increasing prevalence of such applications, it is necessary to understand the properties of these networks as these have implications for both individual incentives and collective welfare.

Previous research has identified several empirical regularities shared by networks in many of these diverse applications. We concentrate on three reasonably robust and prominent empirical characteristics of large networks that have been observed in a variety of settings.

- 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, as we shall discuss).
- Networks tend to have high clustering coefficients relative to what would emerge if the network's links were simply determined by a completely random process.
- The degree distribution of networks tends to exhibit "fat tails" and often approximate a "scale-free" or "power-law" distribution. 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.

While these three characteristics of a network are far from enough to completely characterize a network, together they give us a great deal of information about network structure. An important reason for trying to understand these characteristics, and how they are determined, is that they are critical in determining the overall performance in a network.

Recent papers have looked at explaining these network characteristics. Two prominent studies are by Watts and Strogatz [30],[1] who focussed on finding networks with the joint characteristics of small diameter and high clustering, and by Barabási and Albert [4], who looked at generating scale-free networks. Watts and Strogatz [30] point out the following: start from some sort of symmetric (regular) network with high clustering, but possibly large diameter. They show by simulation that it takes only some minimal random rewiring of a relatively few links to greatly decrease the diameter of the network. Barabási and Albert [4] come from a different angle. It has been known for almost a half century that in a growing society, if individual object size (say degree) grows according to a lognormal distribution over time, and subject to some bounds on object size (e.g., every node forms at least one link upon birth), then the overall distribution of object size in the population will have a scale-free distribution. Barabási and Albert [4] look at a model where new nodes are born each period and choose to form links to existing nodes with a probability proportional to the existing node's degree. This form of preferential attachment generates something like the lognormal growth in the degree of existing nodes, and thus generates a scale free distribution.[2]

While the work to date has made important progress in helping us to understand some of the empirical regularities of large networks, it leaves two

*A full version of the paper appears as Jackson and Rogers [18]. 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. Financial support under NSF grant SES-0316493 is gratefully acknowledged, as is support from the Lee Center for Advanced Networking.

[1] See Watts [29] for a more detailed treatment, and background on high clustering.

[2] There are 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 Kumar et al [20], whose copying method is akin to preferential attachment, as well as Dorogovtsev and Mendes Levene [10], Levene et al [21], Pennock et al [26], and Cooper and Frieze [9]).

very large holes in our understanding.

The first hole is that neither of the approaches for generating networks described above manages to fit all three of the stylized facts. The rewiring approach of Watts and Strogatz ends up exhibiting a small diameter and high clustering, but generally does not exhibit a degree distribution that is scale-free. The preferential attachment based random network generation model of Barabási and Albert [4] (and variations) generate scale-free networks, but those networks have clustering coefficients that are similar to a purely random graph - vanishing in large networks. Thus, neither of these methods of generating networks can be the one underlying most of the large networks that we actually observe.

The second hole is that the processes for generating networks are attempts at answering the question of "how" but not of "why". That is, they are not models of an actual formation process where actors are making some explicit, and say rational, decisions about how to connect the network. They are instead simple models of specific stochastic processes of wiring or rewiring a network that will exhibit some of the desired characteristics. For instance, in the Watts and Strogatz treatment, why should a network look like something that started as a regular network and then was rewired? And with regards to Barabási and Albert [4], why should a link be formed according to preferential attachment?[3] If a node is valued for its degree, then new links would be attached to nodes with maximal degree as opposed to forming links with a chance proportional to their degree. If instead, a node is not valued for its degree why would we see preferential attachment at all?

What we do in this paper is fill both of these holes. We present a process that will exhibit all three of the stylized facts. Moreover, this is a process by which rational economic actors (nodes) search for other nodes and choose the links that they form so as to maximize their economic benefit or utility. The understanding of why the three stylized facts emerge is roughly as follows. The search process occurs through the network itself. This means that nodes with high degree are more likely to be found through the search process. This leads to attachment that has characteristics similar to preferential attachment, which in turn leads to a scale-free distribution of degrees on the higher end. Second, the starting point of the search and/or the cost structure can lead to a higher tendency to form local links. This naturally leads to high clustering. Third, the relatively small diameter comes from the tendency for many nodes to find the same ones to link to (as they are more likely to find and link to nodes which already have large number of links), which generates a diameter which is smaller than the order of a purely random network.

We should mention that an interesting by-product of our analysis is that the degree distribution we obtain approximates a scale free distribution only for the upper tail - that is, only for nodes with relatively high degree, but not for ones with relatively low degree. As pointed out by Pennock et al [26], many of the internet based data sets that are said to be "scale-free," are only scale- free for large degree nodes.

## 2 A Search Model

We propose an 'economic' model of network growth. The agents, who comprise or control the nodes of a network, behave strategically. As they come in contact with other agents, they choose whether to form or not form links with them to maximize their utility.

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.

For any node $n \in N$, let $d_i^{out}(g) = |\{j \in N : g_{ij} = 1\}|$ denote the out-degree of $i$, and $d_i^{in}(g) = |\{j \in N : g_{ji} = 1\}|$ denote the in-degree of $i$. In the case of a non-directed network, these are the same measure, and are simply denoted $d_i$.

The basic model (see Jackson and Rogers [18] for extensions) is based on an explicit search process. 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 (say, 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

[3] There are some papers that give one explanation as to a "why" behind power laws. This is the idea of "HOT" (highly optimized tolerance) systems that underlies Carlson and Doyle [7] and Fabrikant, Koutsoupias, and Papadimitriou [13]. That important idea addresses systems that are optimized, rather than self-organizing. As such the explanation is quite different in both application (for instance, understanding connections among some routers) and approach, and thus quite complementary to that proposed here. Also, such models focus on the scale-free aspect and robustness of the networks (e.g., see Li et al [22]) and not issues of clustering.

the cost. For now, let us assume that the benefit less the cost is independently and identically distributed across $t, t'$ pairs. Jackson and Rogers [18] has other formulations of utility that allow for indirect benefits and externalities from the network structure below. Let $p_r$ denote that the probability.

In addition, (regardless of whether the node forms a link to the parent) the node $t$ searches in 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 - say searching along paths of length at most $k$ from the parent node. Let $p_s$ denote the probability, that the new node attaches to any given one of these nodes found through search. Generally, it is reasonable to have $p_r = p_s$, but we allow for the additional heterogeneity so that we can nest other models (e.g., Barabási and Albert [4]) as special cases.

An expression for the probability that a given existing node $i$ with degree $d_i(t)$ gets a new attachment (in period $t+1$) is roughly[4]

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

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

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

The first expression in (2) is the probability that the node is chosen at random as a parent of the new node and is linked to in that capacity. As there are $t$ existing nodes, and a new node picks $m_r$ of them at random, this probability is self-explanatory. 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 $d_i^{in}(t)$ 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 $\frac{m_r d_i^{in}(t)}{t}$, and then the corresponding probability that the node is identified out of the search through the neighborhoods of the parents is $\frac{m_s}{m_r m}$.

## 2.1 Clustering

We begin the analysis of this model by looking at the clustering coefficient. Given the directed nature of the network, there are various ways in which one might measure clustering. Here, we look at any two links out from a given node, and ask what the probability that those two nodes are linked. So, for instance, if a given web page is linked to two others, then what is the probability that they are linked to each other (one way or the other). Thus the clustering coefficient can be calculated as

$$C(g) = \frac{\sum_i \sum_{k \neq j} g_{ij} g_{ik} \max[g_{jk}, g_{kj}]}{\sum_i \sum_{j \neq k} g_{ij} g_{ik}}.$$

We provide the clustering coefficient for the case where $p_r = p_s = 1$, and when the search by a new node is evenly distributed over each parent's neighborhood. That is, $m_s$ is a multiple $k$ of $m_r$ and $k$ new nodes are picked in each parent's neighborhood.

Proposition 1 *In the search model with $p_r = p_s = 1$, $C(g(t))$ tends to $\frac{2m_s}{m(m-1) - m_r k(k-1)}$ (in probability).*

For the proof, see Jackson and Rogers [18]. If we allow for $p_s$ and $p_r$ to be less than one, then the clustering coefficient is a more complicated expression, but is on the same order, and most importantly, it is bounded away from 0 as $t$ grows.

The fact that the limiting clustering coefficient does not vanish here comes from the search part 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. Most importantly, this distinguishes the search-based model from random graph models, preferential attachment models, as well as the hybrid random graph and preferential attachment models, where the clustering coefficients tend to 0 as $t \to \infty$ (e.g., see Fronczak, Fronczak, and Holyst [16]). Previous models that have been shown to generate high clustering either start from some lattice structure and then rewire, as in Watts and Strogatz [30], or involve some hierarchical structure (see Eiron and McCurley [11]). Substantial evidence suggests that large networks indeed exhibit clustering measures much larger than would be predicted by either purely random processes or models based on preferential attachment, as well as the hybrid versions.[5]

[4] 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 the node is found by more than one method of search. Nevertheless, is a very accurate approximation for when the network is large (i.e., $t$ is large) relative to $m_r$ and $m_s$, as these adjustments vanish.

[5] For instance, Watts [29] gives a clustering coefficient of 0.79 for the network consisting of movie actors linked by movies in which they have co-starred. Networks of researchers linked by co-authored papers have also been analyzed in various fields of study. For instance, Newman [24] gives clustering coeffi-

## 2.2 A Mean Field Analysis of the Degree Distribution

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

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

If we start the system with each node $t$ having in degree counted as $d_0$ (for instance 0),[6] when it is born at time $t$, then we can solve the differential equation given by (3) to find

$$d_i^{in}(t) = (d_0 + rm)\left(\frac{t}{i}\right)^{\frac{p_s m_s}{m}} - rm,$$

where $r = \frac{p_r m_r}{p_s m_s}$ is the relative fraction of links that are formed at random compared to through the search.[7]

Proposition **2** *The degree distribution of the above process has a cumulative distribution function of*

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

*for $d \geq d_0$.*

The proof of Proposition 2 appears in Jackson and Rogers [18], and is similar to results found in many papers that have used mean field approximations to estimate large random network properties.

As a check on the mean-field approximations, where we match the analytic solution from Proposition 2 with simulations of random process itself. The two match up well for all degrees, and for a variety of different parameters that we have run. See Jackson and Rogers [18] for details.

Figure 1 illustrates these effects by showing the complimentary cdf $1 - F(d)$ of the degree distribution for three parameterizations of the model. The left panel shows 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. [Let us point out that in a log log plot of the complementary cdf, the larger part of the graph comprising the right hand side can actually only be due to a small fraction of the data. We discuss this in more detail in Jackson and Rogers [18].] 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.

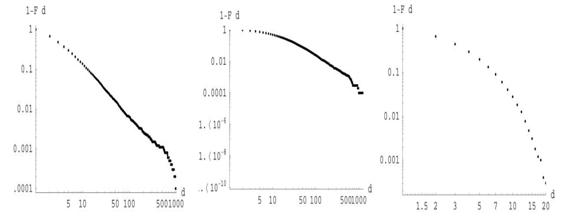

**Figure 1.** *Simulations are based on $T = 10,000$ periods. (left) Entering nodes connect to two parents as well as to two nodes from each of the parents' outlinks ($m_r = 2$, $p_r = 1$, $k = 2$, and $p_s = 1$). (middle) Entering nodes connect to nine parents and to one node from each of the parents' outlinks ($m_r = 9$, $p_r = 1$, $k = 1$, and $p_s = 1$). (right) A purely random graph where entering nodes connect to two parents and no other nodes ($m_r = 2$, $p_r = 1$, and $m_s = 0$).*

Compare these distributions with those in Figure 2, which contain data from coauthorship networks (left) from Newman [24] and the world wide web from Albert, Jeong, Barabási [2]. While the latter appears to exhibit a scale-free distribution, the former clearly does not. The search model accommodates both cases, and suggests that the role of search is more prevalent in the www case than for the coauthor network.

## 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 [12]. For some special cases we can deduce limits on the diameter by piggy-backing on some powerful results due to Bollobás and Riordan [6].

cients of 0.496 for computer science, and 0.43 for physics, while [Grossman] gives a measure of 0.15 in mathematics. Several authors have also analyzed clustering in the world wide web. For instance, Adamic [1] gives a 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).

[6] 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.

[7] This presumes that $p_s m_s > 0$, as otherwise (3) simplifies and has a different solution, as discussed in Jackson and Rogers [18].

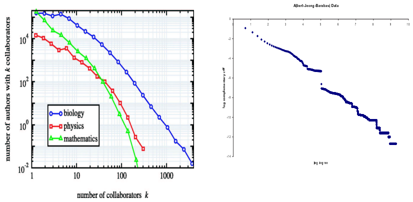

**Figure 2.** *(left) Data from Newman [24] containing the frequencies of authors with varying numbers of coauthors, which are clearly not scale free. (right) Data from Albert, Jeong, Barabási [2] showing the complimentary cdf of web-page degrees, which is approximately scale-free.*

Proposition **3** *If $p_r = 0$, $p_s = 1$, $m_s = 1$, and $m_r \geq 2$, then the resulting network will consist of a singlecomponent with diameter*[8] *proportional to $\frac{\log(t)}{\log\log(t)}$, almost surely.*

The proof follows from Bollobás and Riordan [6].[9]

We conjecture that increasing the parameters $p_r$ and $m_s$ and decreasing $p_s$ (provided $m_r \geq 2$) should not affect these results, and this is confirmed by following the heuristic test suggested on page 24 of [6]. It is worth noting that the constraint that $m_r \geq 2$ is critical. Intuitively, it is this independent search that allows a node to form a bridge between different existing neighborhoods of the network, thus reducing path length. 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, then 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. In the simulations, we calculate a crude upper bound on the diameter of a network by simply doubling the size of neighborhood a given node much search out in order to see the entire population. For the simulations presented in Figure 1 (middle), using the node with highest degree, this crude upper bound is typically $\bar{\delta} = 6$ for $T = 10,000$. Given that $\ln(T) = 9.21$ and $\frac{\ln(T)}{\ln\ln(T)} = 4.15$, and that the upper bound we calculate is not tight, this suggests that the diameter for the search model is indeed of order smaller than $\ln(T)$.

[8] 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.

[9] We need to allow nodes to self-connect and enter with degree 1, in order to apply their proof.

# References

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

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

[3] Barabási, A. (2002), *Linked*, .

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

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

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

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

[8] 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).*

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

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

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

[12] Erdös, P. and A. Rényi (1960) " ," Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 5,

[13] Fabrikant, A., E. Koutsoupias, and C. Papadimitriou (2002), "Heuristically Optimized Trade-offs: A new paradigm for power laws in the Internet," *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming.*

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

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

[16] 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.

[17] Jackson, M.O. (2004) "A Survey of Models of Network Formation: Stability and Efficiency," forthcoming in *Group Formation in Economics; Networks, Clubs and Coalitions* , edited by G. Demange and M. Wooders, Cambridge University Press: Cambridge U.K., http://www.hss.caltech.edu/ ∼ jacksonm/netsurv.pdf.

[18] Jackson, M.O. and B. Rogers (2004) "The Strategic Formation of Large Networks: When and Why do We See Power Laws and Small Worlds?," preprint: Caltech, http://www.hss.caltech.edu/ ∼ jacksonm/netpower.pdf.

[19] 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.

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

[21] 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.

[22] 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.

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

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

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

[26] 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.

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

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

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

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

[31] 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.

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